TECHNISCHE FACHTAGUNG JUNI 2006

System x & BladeCenter Performance Update

Anastasios Panagou
Advisory IT Specialist – FTSS System x & BladeCenter
IBM Schweiz
apa@ch.ibm.com

IBM Systems

Agenda

- **Performance**
  - **Benchmark Types**
- **Network Subsystem Performance**
  - **TOE, IOAT and 10Gbit Ethernet Throughput**
- **Memory Subsystem**
- **Intel / AMD**
  - **CPU's**
  - **IBM EXA-3 Technology**
- **Hard Disks**
- **IBM Systems**

Performance & Benchmarks

Benchmark	Type of workload
TPC-C and TPC-E	Online transaction processing
TPC-H	Ad-hoc decision support
TPC-W and TPC-App	Transactional Web e-Commerce
SPECweb	Web serving static and dynamic pages
SPECweb_SSL	Web serving with Secure Sockets Layer (SSL)
SPECjbb	Server-side Java™
SPECjAppServer	J2EE-based application server
SPEC HPC	High performance computing, CPU, interconnect, compiler and I/O subsystems
SPEC CPU	Compute intensive, integer and floating point performance
Oracle Applications	Models the most common transactions on the seven most used Oracle Application modules
SAP Standard Application	Suite of benchmarks for mySAP Business Suite.
BaanERP	Transaction processing environment of iBaan ERP applications
Notesbench	Simulates Domino® workstation-to-server or server-to-server operations
Exchange MAPI Messaging	Measures the maximum messaging throughput of a Microsoft® Exchange Server
Linpack HPL	Solving a dense system of linear equations Used to compile the top 500 supercomputer list.

http://www.redbooks.ibm.com/abstracts/redp3957.html

TPC Transaction Processing Performance Council http://www.tpc.org/

spec http://www.spec.org/

Benchmark Types

Online Transaction Processing (OLTP)

1. Memory subsystem
2. Disk subsystem
3. CPU subsystem
4. Network subsystem

TPC-C
TPC-E

Decision Support (DSS)

1. Disk subsystem
2. CPU subsystem
3. Memory subsystem

TPC-H

Web servers

1. Memory subsystem
2. CPU subsystem
3. Network
4. Disk subsystem

TPC-W
SPECweb

Exchange MAPI
Messaging

Mail and collaboration

1. Disk subsystem
2. Processor subsystem
3. Memory subsystem

SAP, Oracle,
Notesbench

Application server and ERP

We have different characters on the workload depending on the design. Generally speaking the middleware is the layer that is normally most constrained in a 3-tier solution because that is where all the applications are.

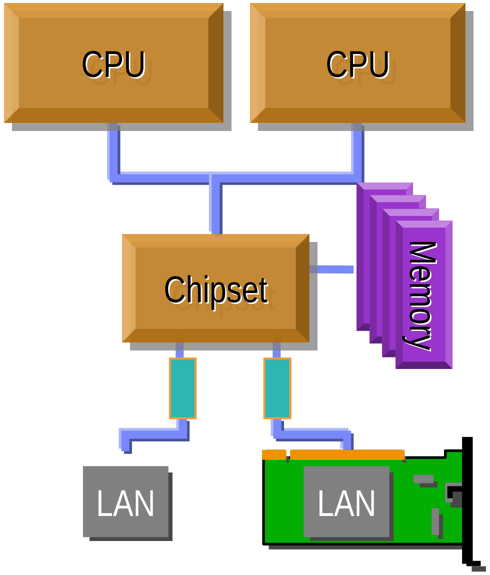
3-tier solutions consists of the following:

- Front end layer
- Middleware application layer
- Back-end database layer

In a 2-tier solution, the middleware and the back-end are integrated as one part. In most cases applications and ERP solutions are designed as 3-tier solutions.

1 = most important, 4 = least important

Network Subsystem Performance Issues

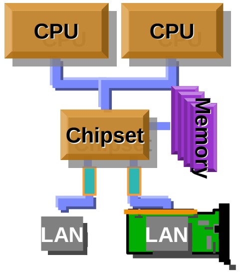


Current TCP/IP Challenges

- **As networks get faster CPU processing power needed to drive the network is becoming a bottleneck**

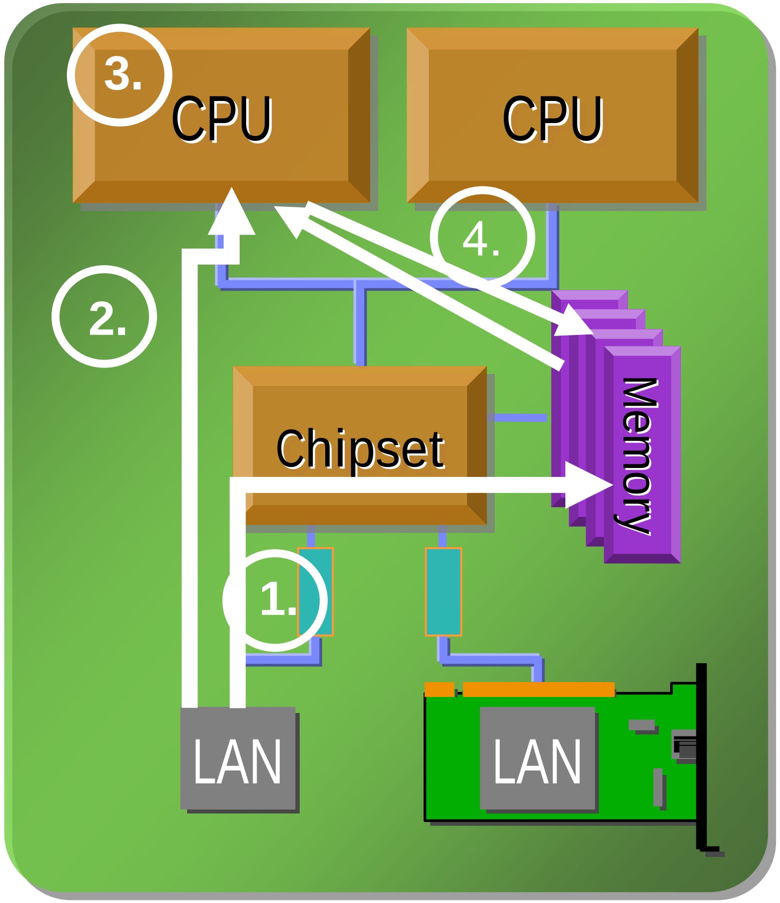
- **Standard NICs have evolved to be more efficient but…**
  - **10Gbit/sec Ethernet running full-speed often requires more processing power than available in most 4-way servers**

- **Two technologies are emerging to address networking overhead**
  - **TCP/IP Off-load Engine (TOE)**
  - **IOAT I/O Acceleration Technology by Intel**



Network Architecture – Standard NIC

1. Packet received by NIC and packet DMAed to driver/kernel buffer in main memory
2. Network controller interrupts CPU to signal receive packet command
3. CPU processes TCPI/P and TCP descriptor and headers
4. Data is copied from kernel memory space to user memory space by the CPU

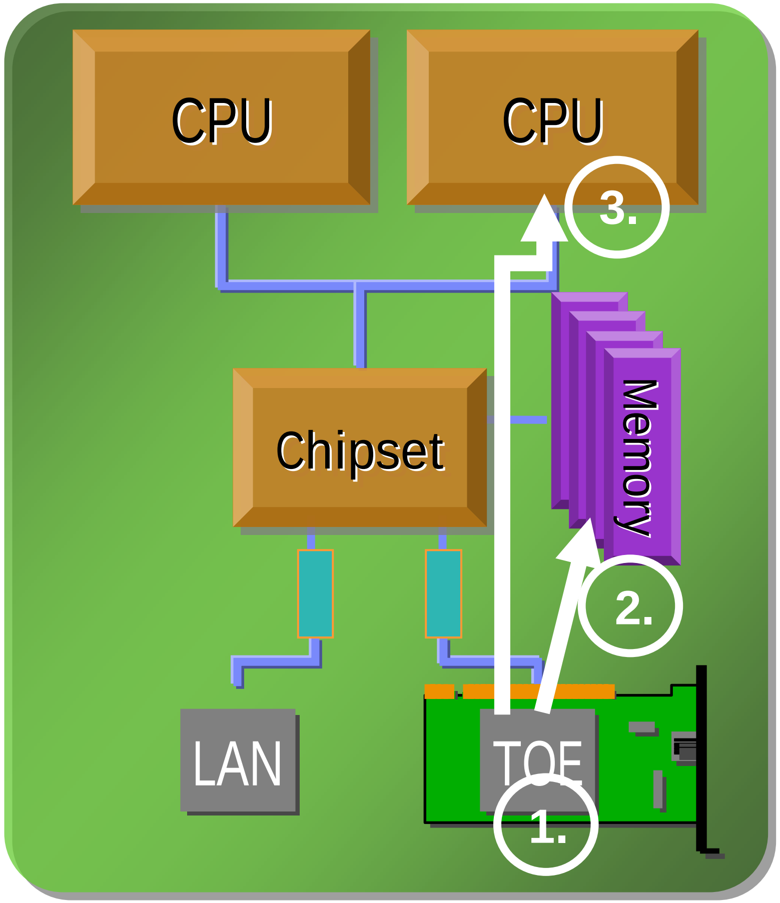


Potential bottlenecks

1) Interrupt Process and Multiple Memory Accesses by the CPU
2) TCP Protocol Processing
3) CPU ⟷ Memory Copies

TCP/IP Offload - TOE

1. **Packet received by TOE processes TCP/IP headers**
2. **Data is DMA-ed to user memory space directly**
3. **Network controller interrupts CPU to signal arrival of packet and location of data in application memory space**

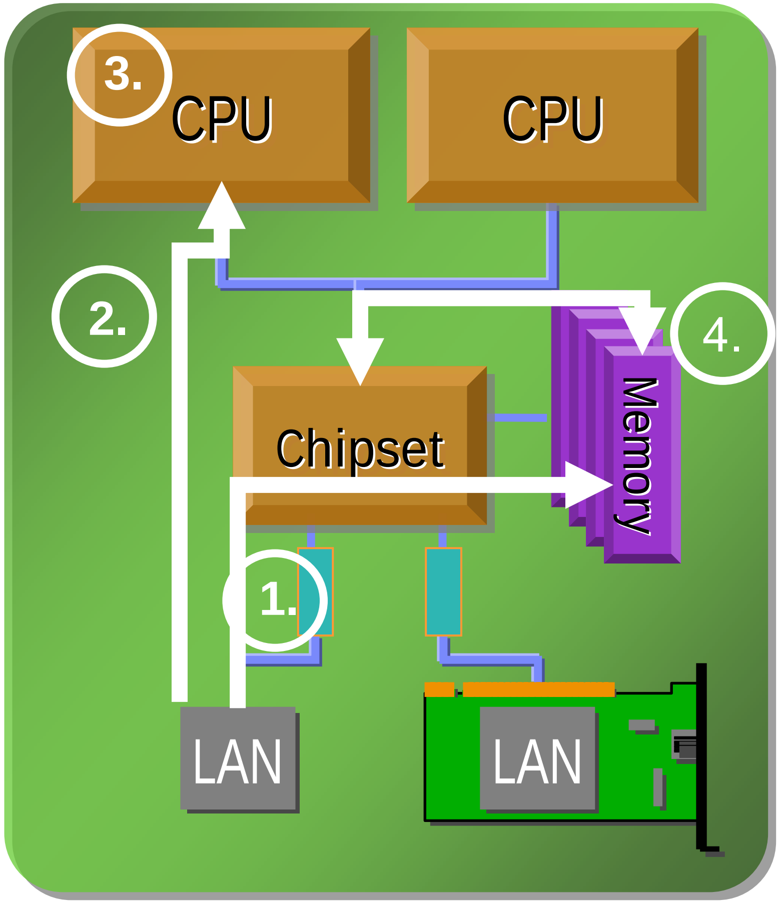


Benefit

- **Less code processing by CPU**
- **Fewer CPU data copies**

I/O Acceleration Technology – Intel (IOAT)

1. **Packet received by NIC and DMAed to driver/kernel memory space**
2. **Network controller interrupts CPU to signal arrival of packet**
3. **CPU processes TCP/IP descriptors and headers**
4. **Data is copied from kernel memory space to user memory space by DMA engine contained in chipset**



Benefit

- **Few if any data copies by CPU**
- **First version will only help receive performance since copies will be done only on frames that are moving from TCP/IP space to application space**

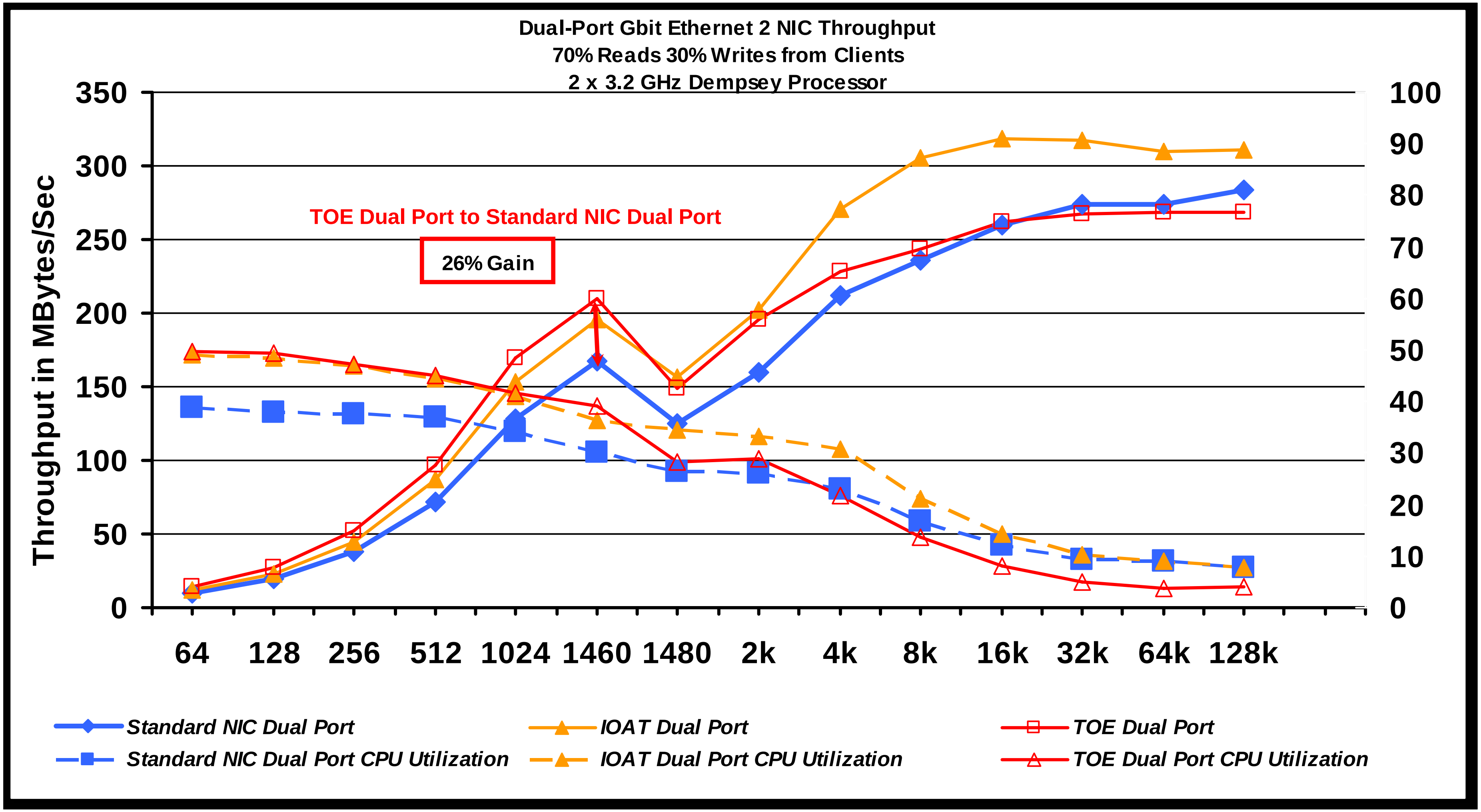
TOE vs. IOAT Comparison

- **TOE employs a NIC based processor and offloads protocol**
  - **Performance or number of connections limited by NIC configuration**
  - **Off-load can be accomplished on any system by installing TOE adapter, driver, and TOE aware OS**
    - **Requires x64 Windows Chimney to obtain consistent Windows interface**
    - **Linux has not declared support for TOE**

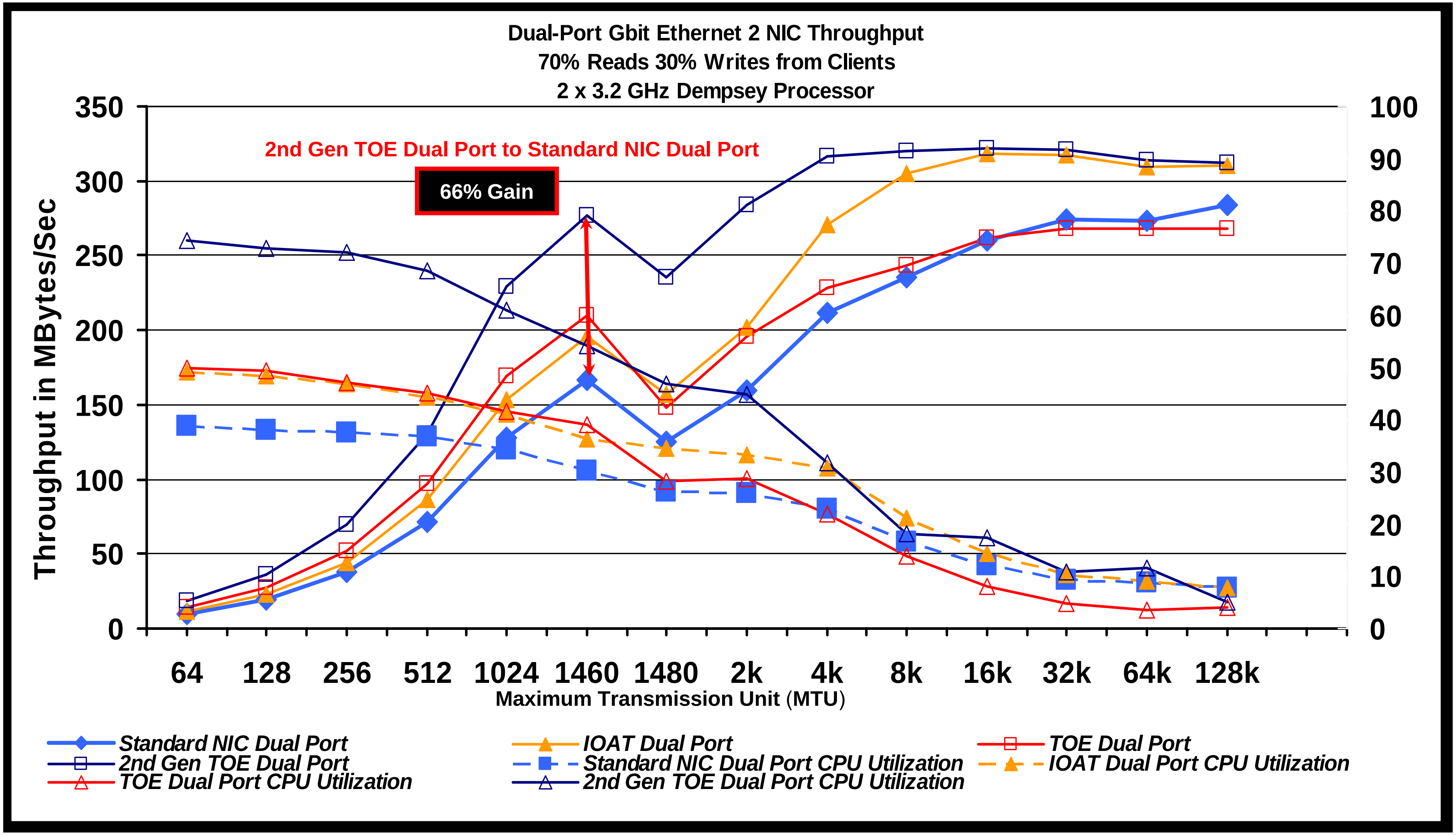
- **IOAT is accomplished by new Intel chipset with enhanced data moving capabilities, new API to control data flow, and driver**
  - **"On-loads" TCPI/P to processor with more efficient copy mechanism**
  - **Initial versions only enhance receive path**
    - **Since most servers send data, initial gains will be moderate**
    - **Send path optimization in future IOAT versions**

Early IOAT vs. Standard NIC Vs. TOE Measurements

PRELIMINARY IOAT & TOE vs. Standard Intel Pro 1000 NIC Windows 2003 SP1 64-bit – Dual NIC



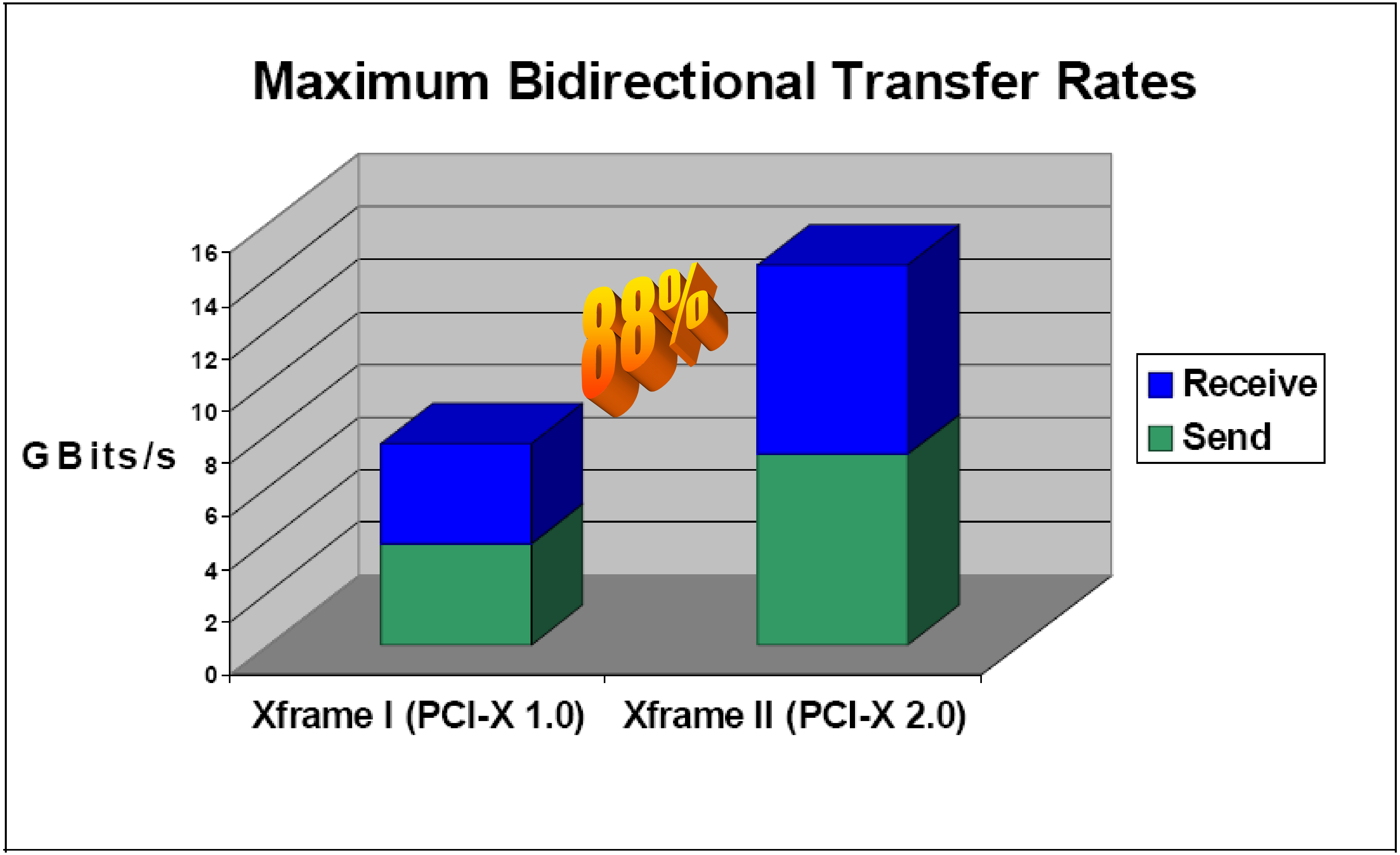
Second Generation TOE – Even Greater Scalability Windows 2003 SP1 64-bit – Dual NIC



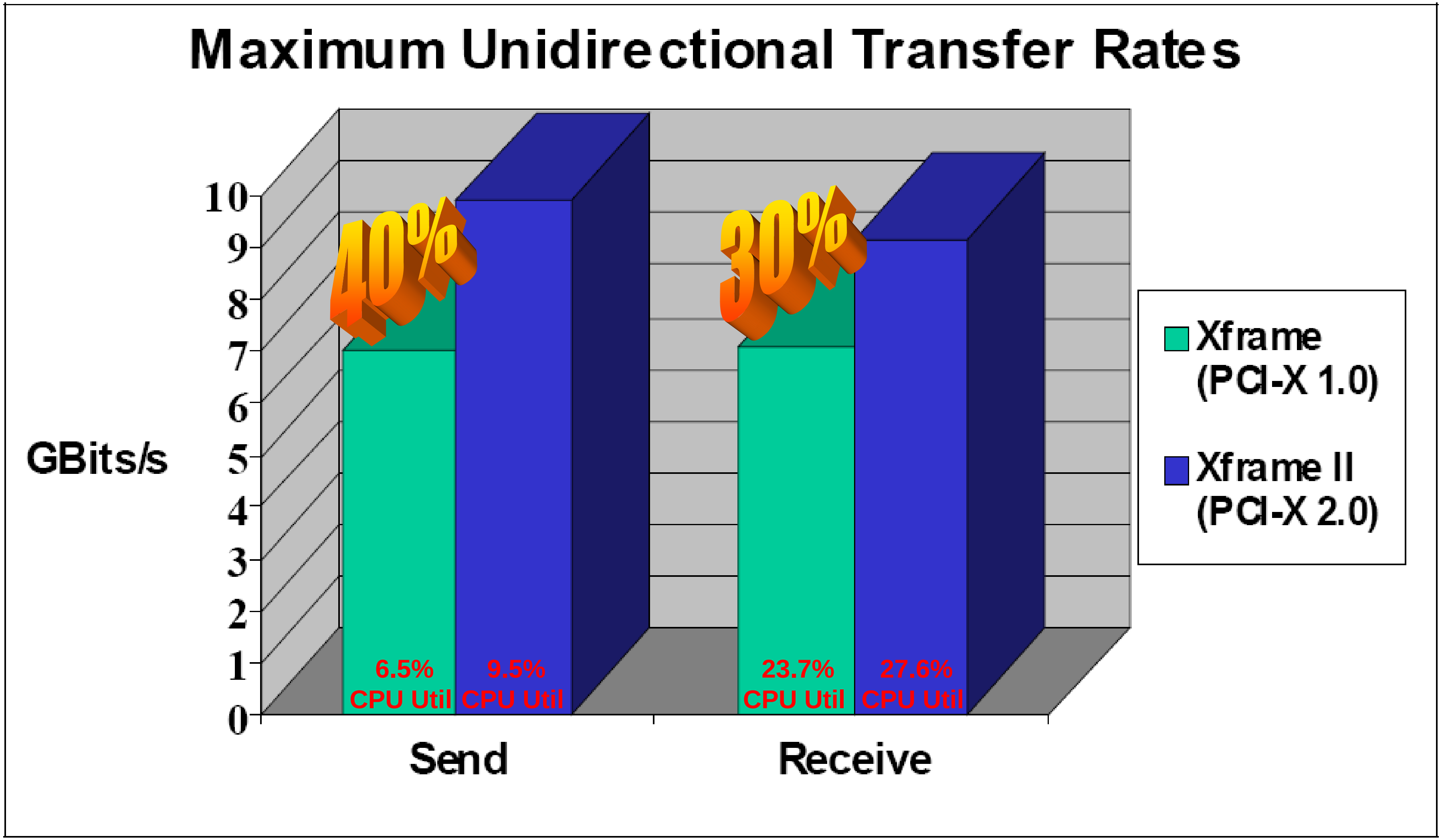
10Gbit/Sec NIC Measurements

10 Gigabit Ethernet Throughput in x260/x366/x460

Maximum throughput achieved with jumbo frames, 512KB transfers



10 Gigabit Ethernet Throughput in x260/x366/x460
Maximum throughput achieved with jumbo frame, 512KB transfers



Summary

- **TOE and IOAT are in the early stages**
  - **TOE is currently supported in Windows 2003 x64 SP1 but not in Linux**
  - **IOAT will require driver changes and new HW but will be supported by both Linux and Windows**

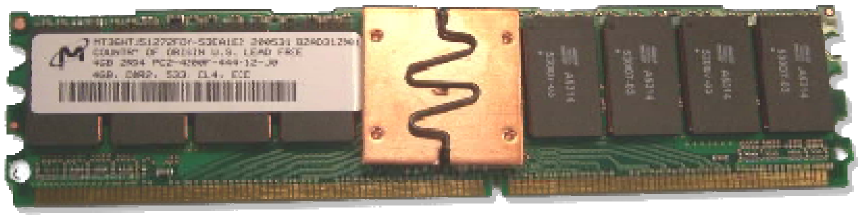
- **Compared to standard NIC all versions of TOE provide good scaling and compelling gains in throughput (we measure peak of 20 – 46% gains) especially for middle range of (128 – 8KB) packet sizes**

- **Current generation TOE provides significant gains (peak of about 29%) with single NIC and about 66% higher throughput when using 2 or more NICs over standard NIC for middle range of packet sizes**
  - **Future versions (in product early 2007) of TOE will provided even greater gains in throughput when PCI-X bridges are removed and native PCI-E silicon is integrated into solutions**

- **For single NIC IOAT gains are not significant, but with dual-NICs we measure 17 – 30% greater throughput and/or reduction in CPU usage**
  - **Greatest gains with multiple NICs and large transfer sizes**

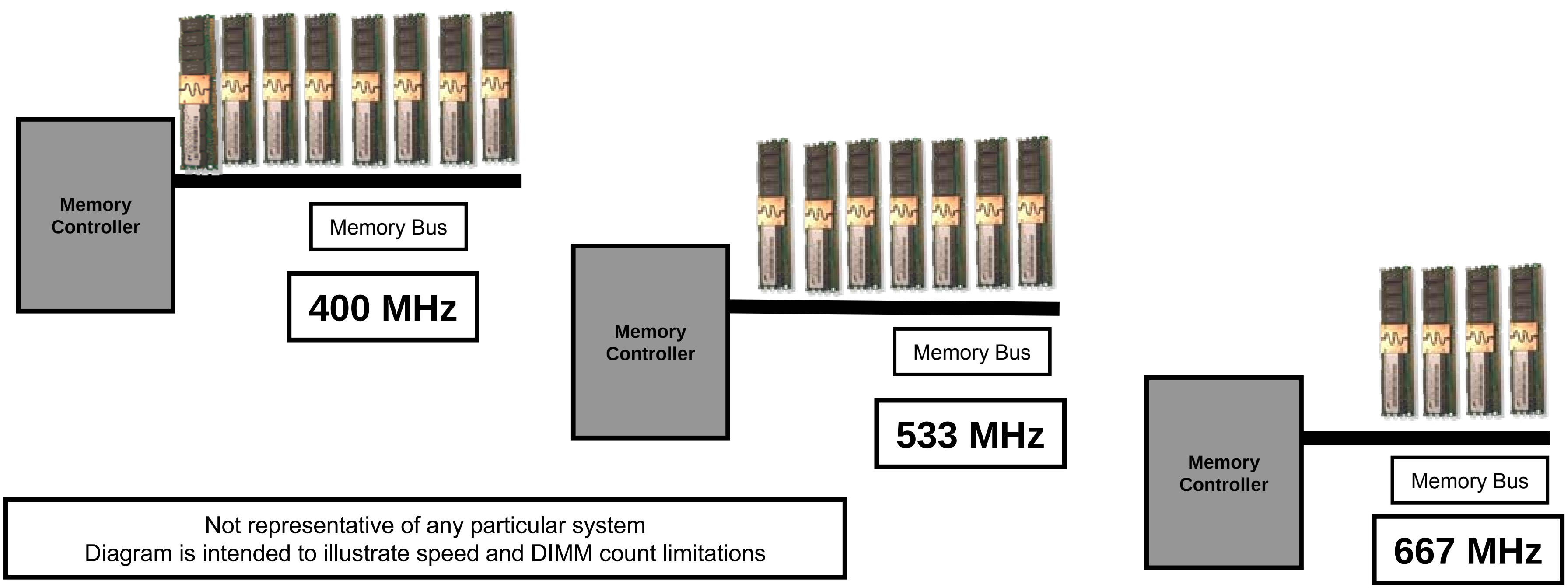
- **Systems based on x3 chipset (x260/x366/x460) capable of full 10Gbit Ethernet throughput**
  - **Lab measurements indicate maximum sustain throughput of about 30Gbit/Sec with multiple 10Gbit Ethernet Adapters and large packet with jumbo frame transactions**
    - **Keep in mind that real-world throughput will greatly depend upon application characteristics**
    - **For 10 Gbit Ethernet Throughput Paper see:**
      - http://www.ibm.com/common/ssi/fcgi-bin/ssialias?infotype=SA&subtype=WH&htmlfid=XSW01607USEN&attachment=XSW01607USEN.PDF&appname=STG_XS_USEN_WH

Memory Subsystem Performance Issues

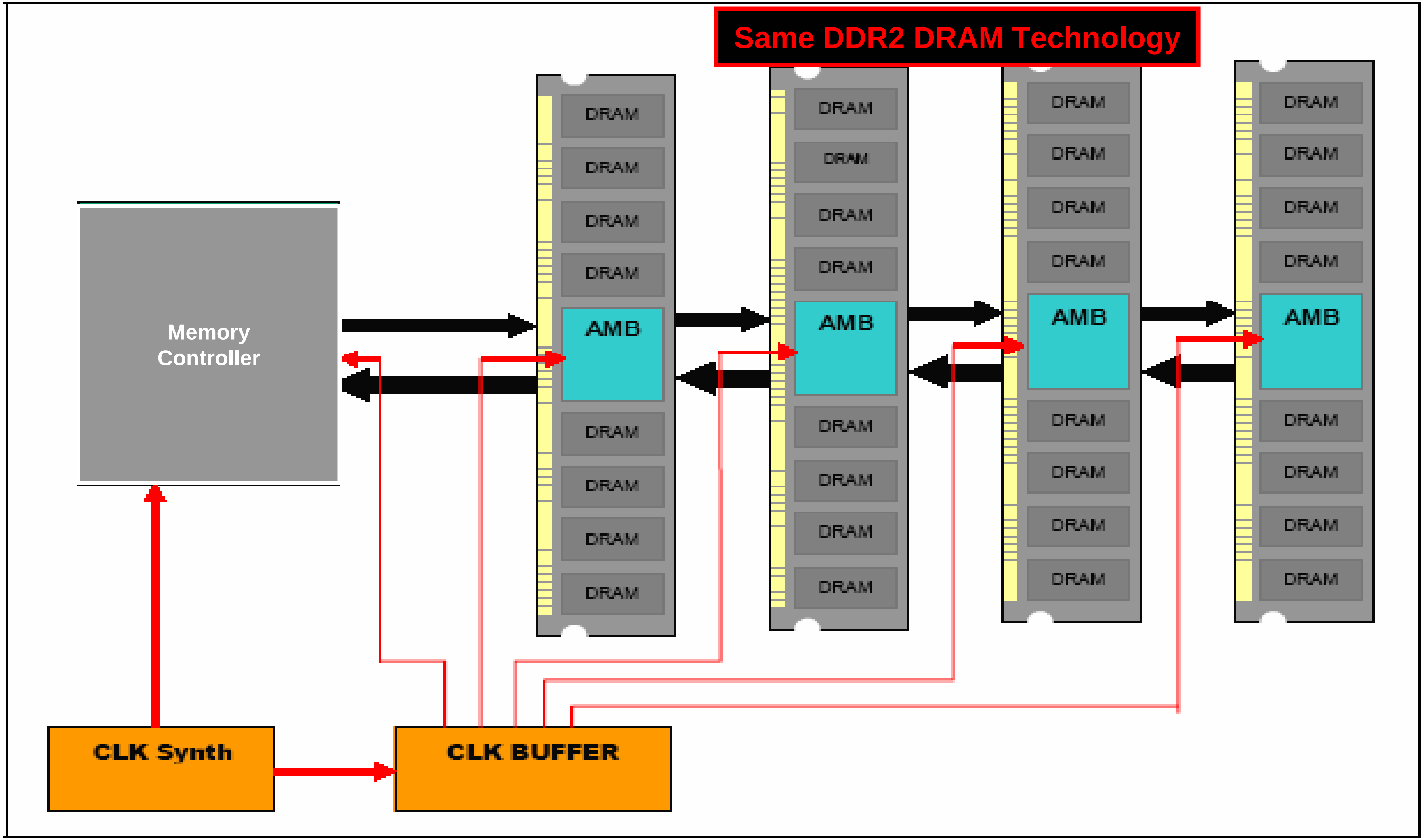


Problem Statement

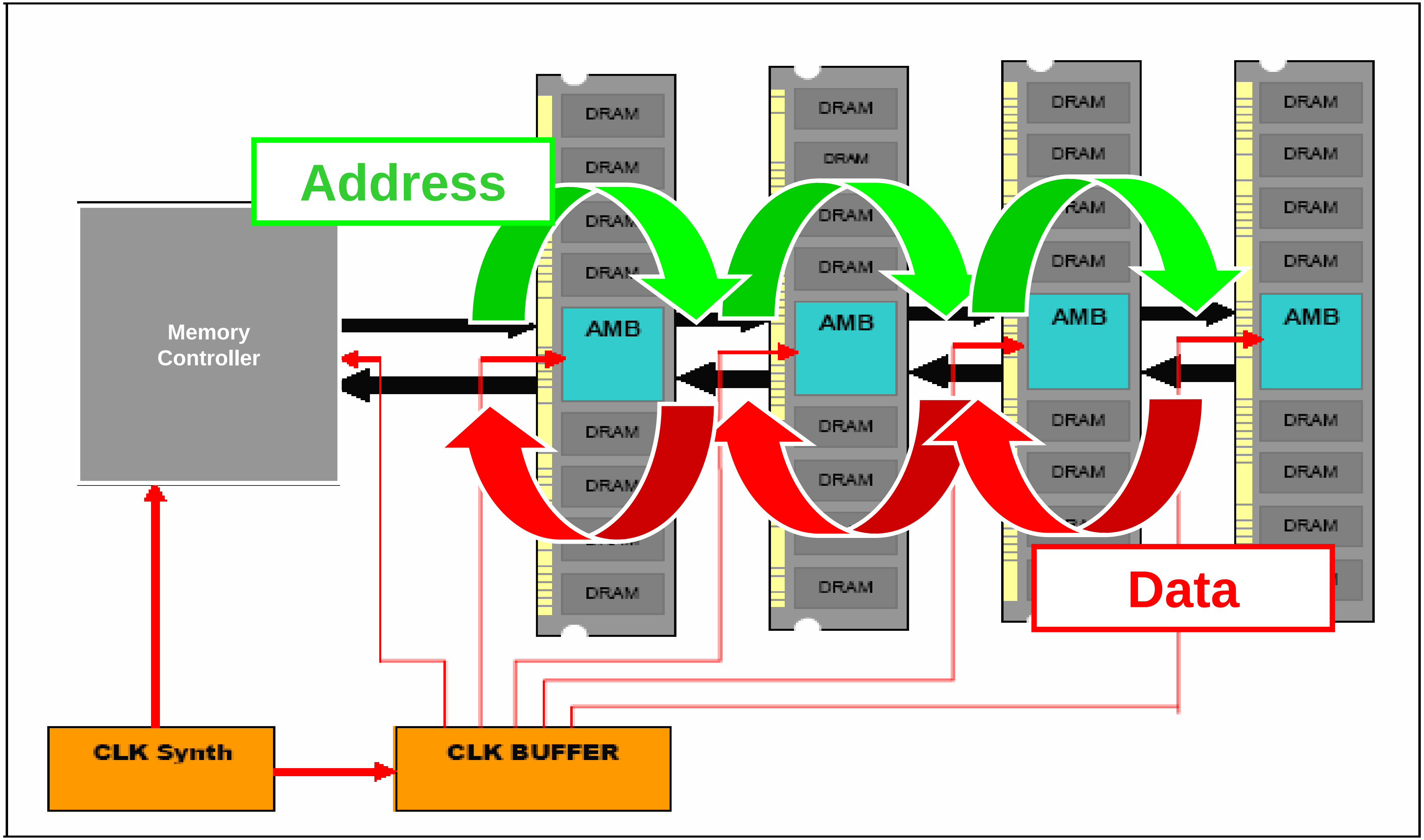
- **DDR2 DIMMs add electrical loading to memory bus**
  - **This means that as memory clock speed increases the number of DIMMs that can be supported on the memory channel decreases because of electrical loading**



FBDIMM Solves Problem With Serial Memory Bus And On-DIMM Advanced Memory Buffer (AMB)



FBDIMM Serial Bus Add Latency Due to Hops



But FBDIMM Serial Interface Reduces Wiring Complexity And Enables Greater Number of Memory Channels

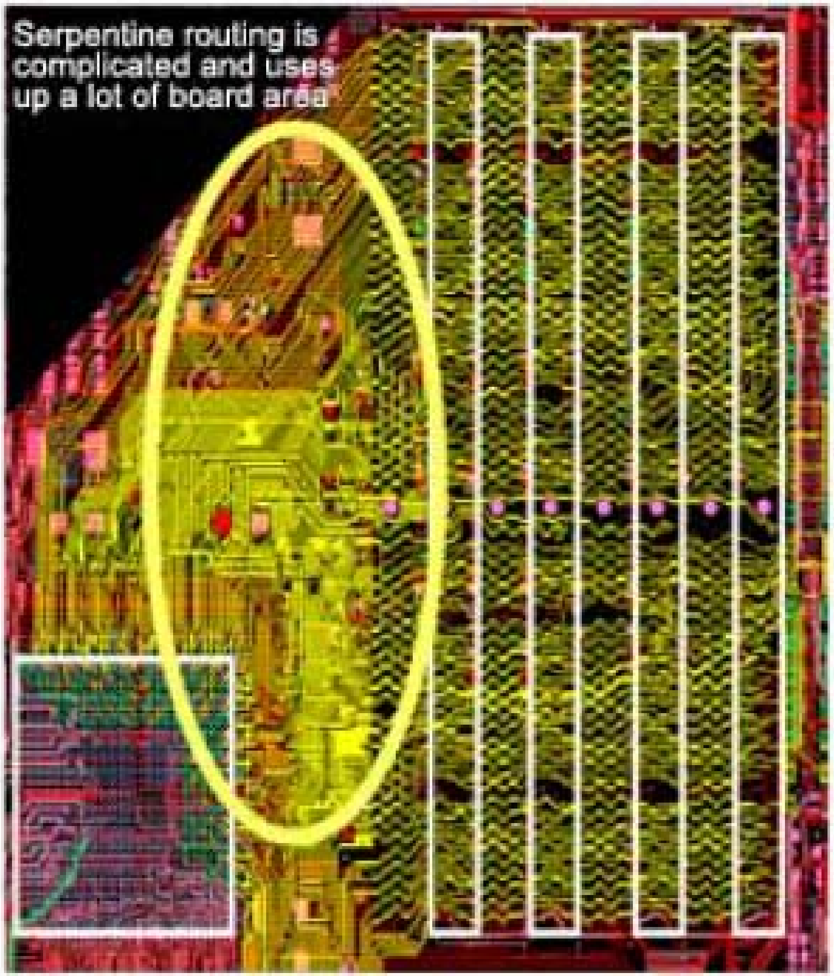


Figure 3. DDR2 Registered DIMMs: 1 Channel, 2 Routing Layers with 3rd layer required for power

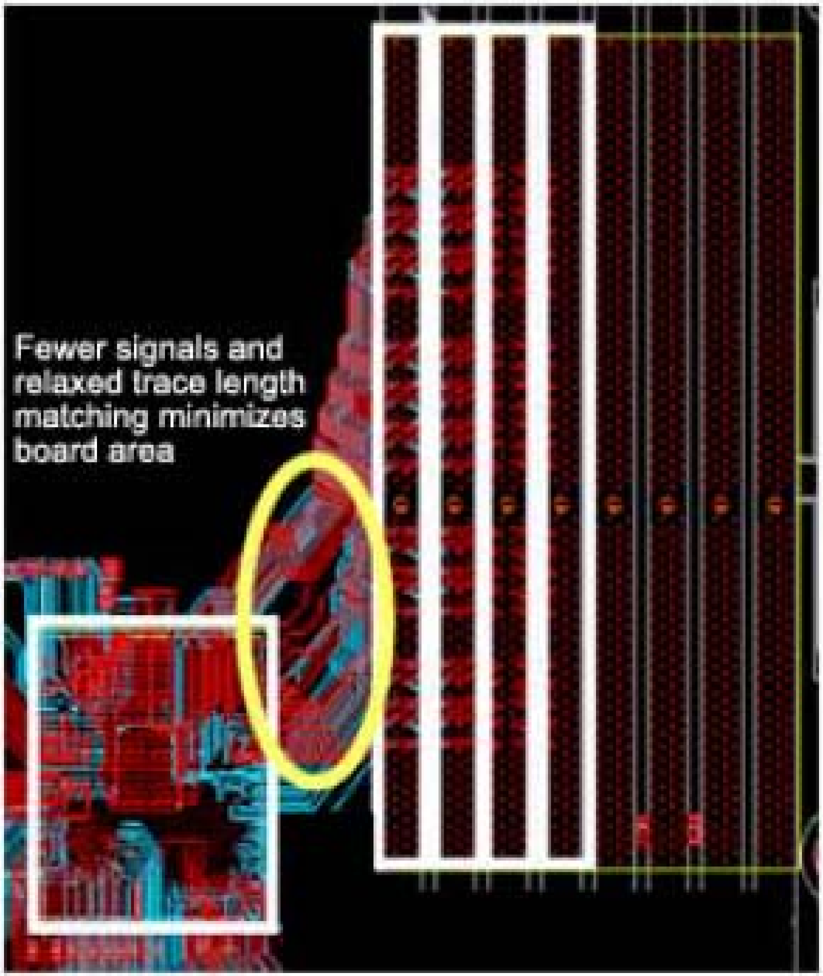
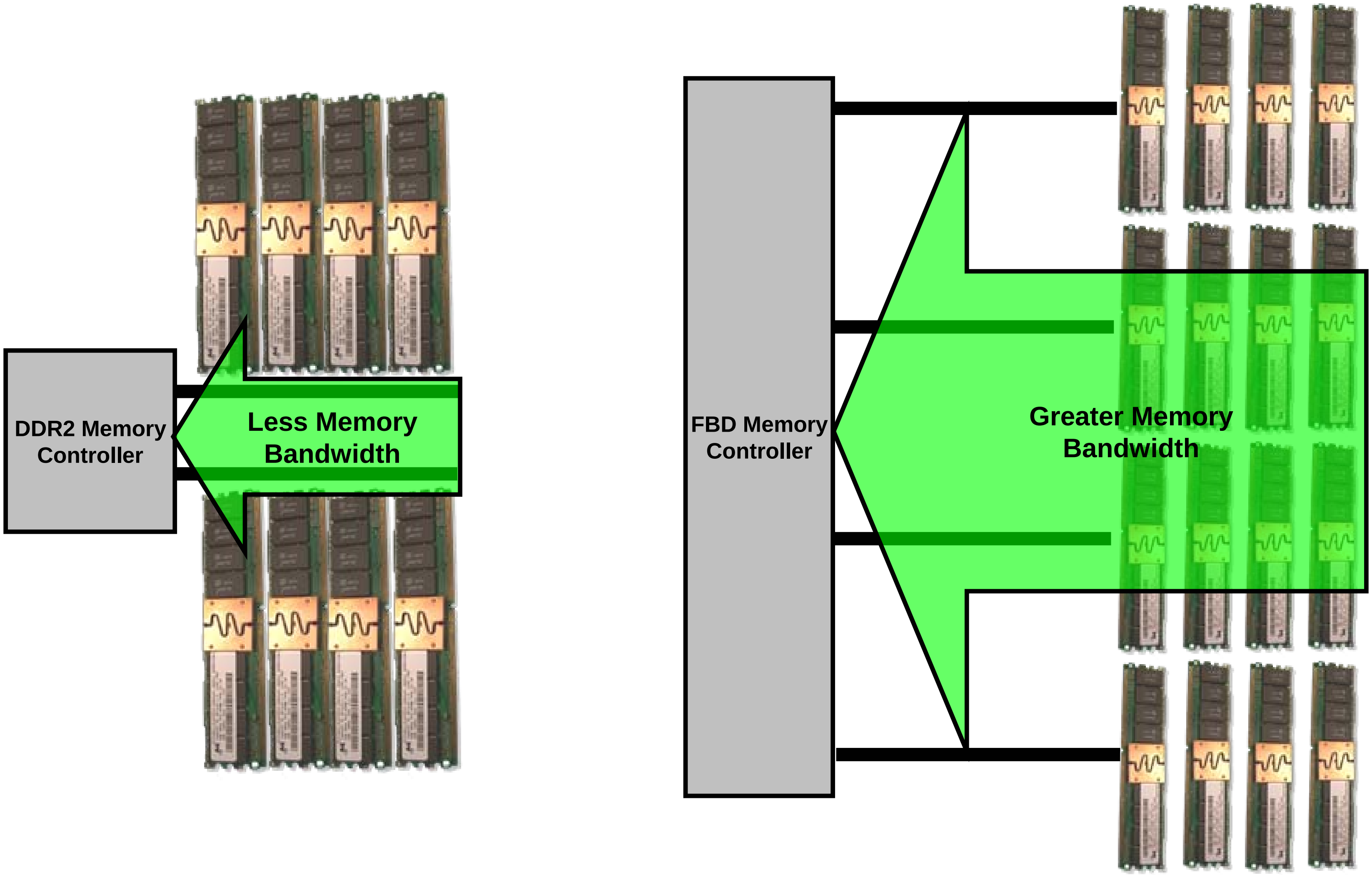
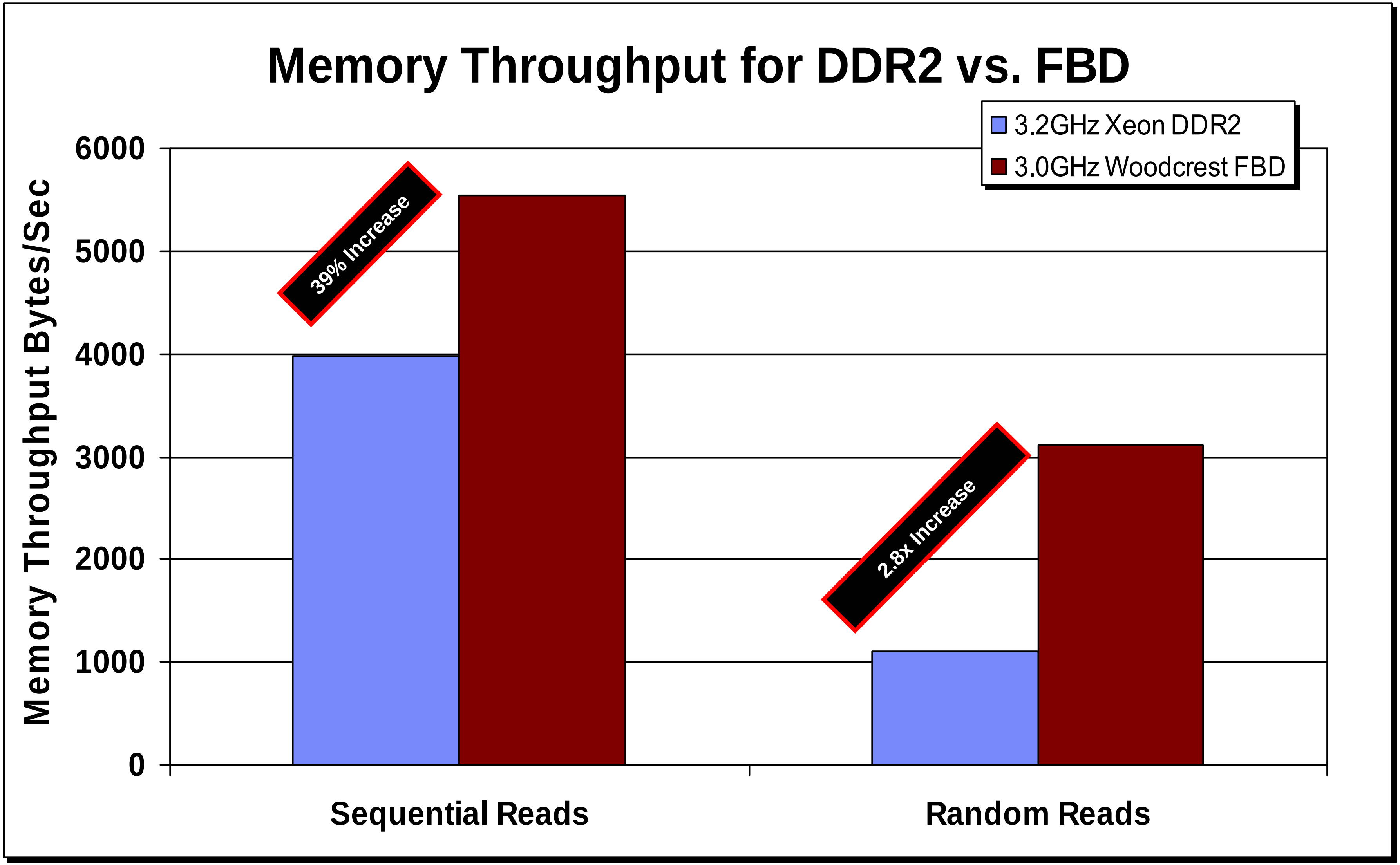


Figure 4. FB-DIMMs: 2 Channels, 2 Routing Layers (includes power delivery)

Additional Memory Channels = Greater Capacity And Greater Throughput Which Offsets Additional Latency Under Load



DDR2 vs. FBD Memory Throughput



Memory Summary

- **Existing DDR2 memory employs multi-drop parallel bus**
  - **Electrical loadings increase as DIMMs are added to the bus**
    - **This limits the speed of the memory bus**
  - **Parallel bus limits number of memory channels in system**
    - **Physical wiring space limits number of memory channels on planar boards**
    - **Memory controller pin-count too great with more than two channels**

- **FBDIMM solves problem by placing an Advanced Memory Buffer (AMB) on DDR2 DIMM and employs a serial memory bus**
  - **Serial bus greatly reduces wiring requirements and enables greater number of memory buses supported in a system**
    - **This increases capacity and throughput**
  - **Serial AMB adds latency and increases DIMM power consumption**
    - **But greater throughput results in LOWER average latency when under load, improving performance**
    - **Second generation AMB will consume even lower power**

Processor Subsystem Performance Issues
Snoop Filters

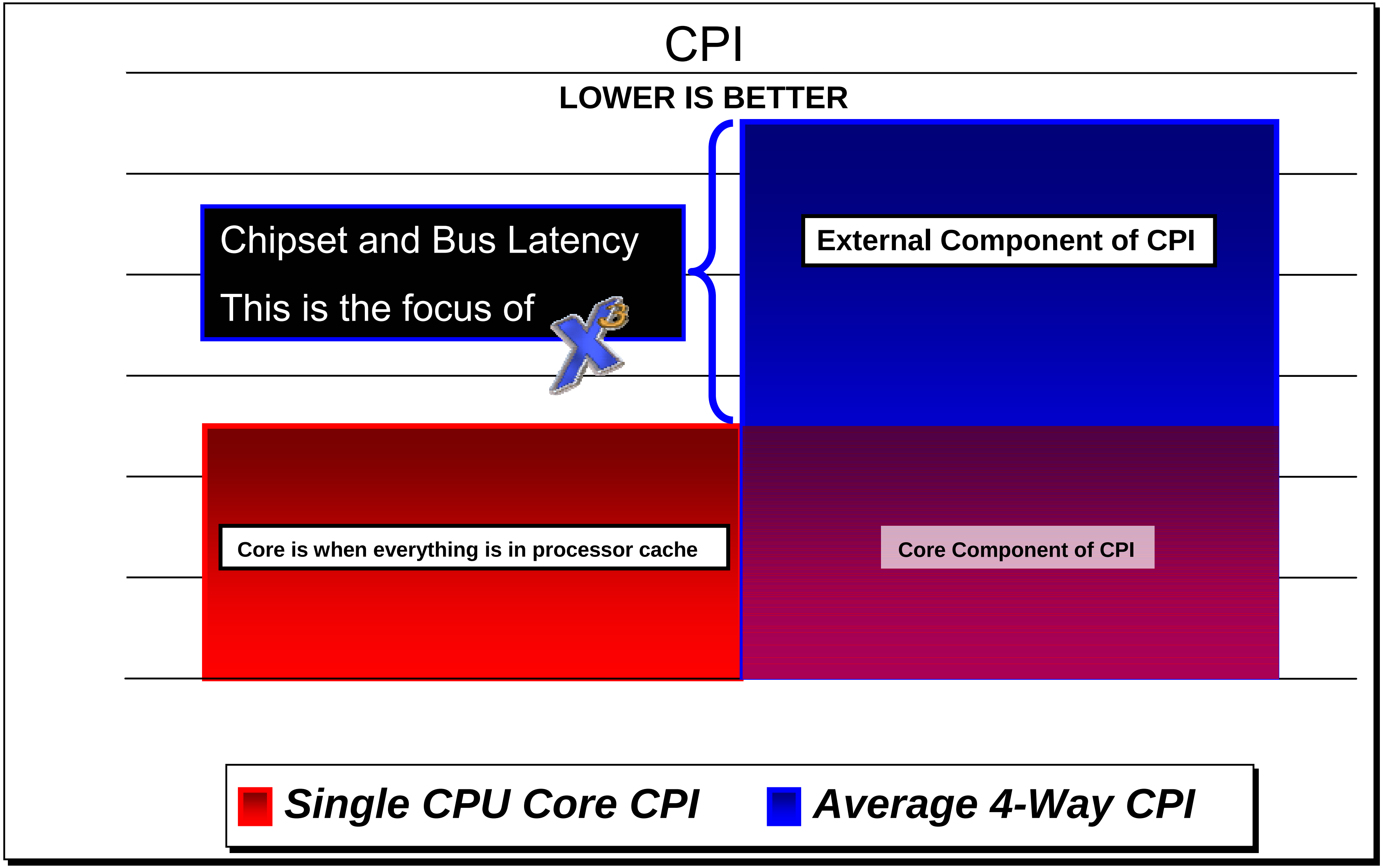
What is CPI?

- **CPI – <u>Clocks Per Instruction</u> is a key metric used by processor and system designers to identify <u>performance efficiency of a processor</u>**
  - **Much like Miles Per Hour or Miles Per Gallon used for automobiles**

- **CPI – Can first be explained by looking at the clocks per instruction for a processor when all instructions and data are in the fastest processor cache**
  - **We call this Infinite Cache CPI or Core CPI**

- **But in a real-world system, not everything comes from fast cache**
  - **Frequent cache misses take much longer to service than a cache hit**
    - **Processor must wait longer period before obtaining data or instructions from memory**
    - **This increases the Average CPI of processor running real applications**

A Closer Look At CPI



The x3 Chipset Focuses On Reducing External CPI (or Latency)

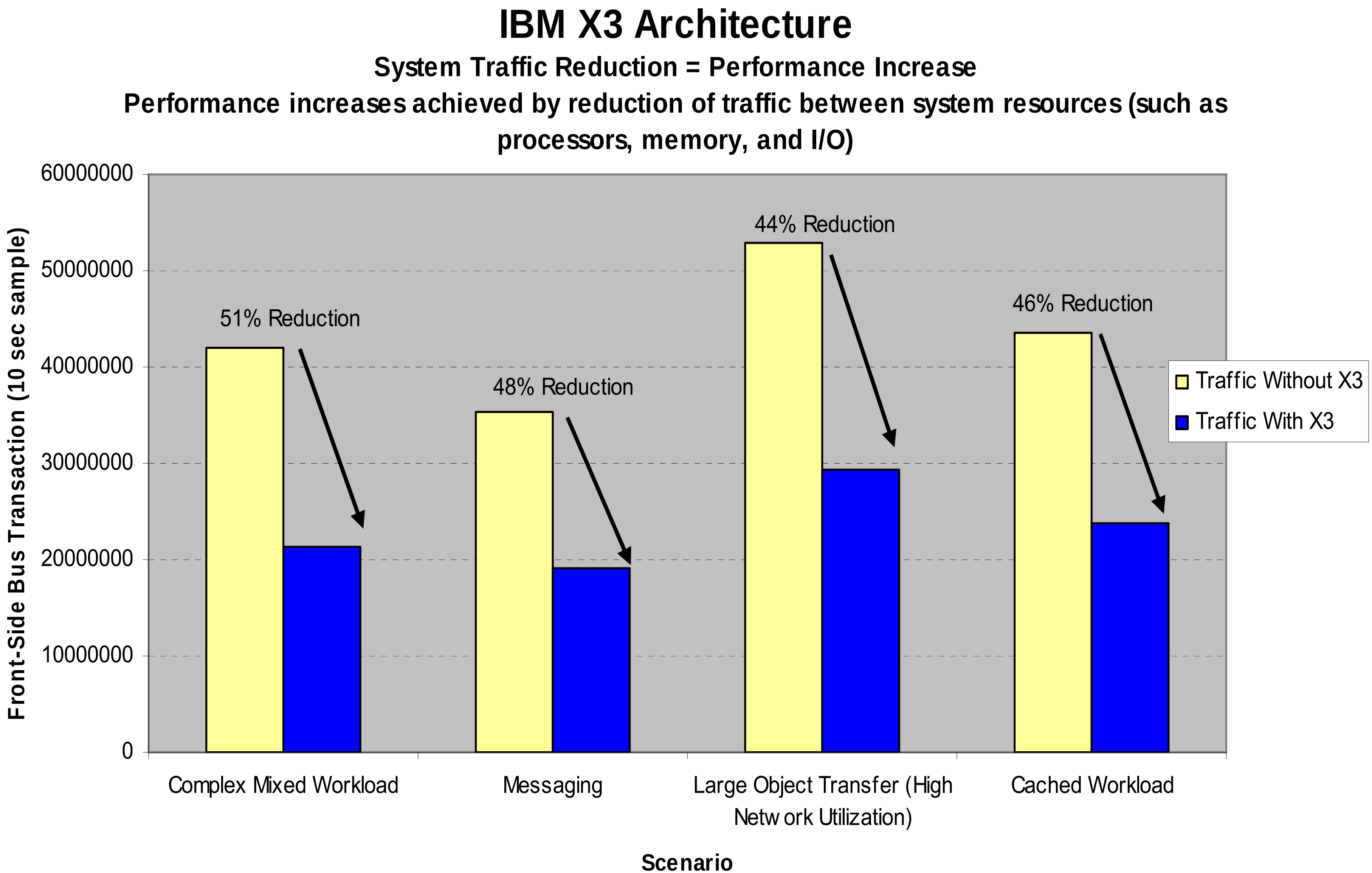
- **Performance can be improved dramatically by reducing the average number of processor clocks needed to process instructions**

- **But core CPI can only be changed by the processor designer?**
  - **One can gain greater performance by improving core performance**
    - **Larger caches**
    - **Faster clock speed**
    - **Optimizing application**
  - **But faster processors with larger caches are expensive**
  - **And this only effects the component of time spent <u>inside</u> the processor**
    - **Which is usually smaller than the total time spent waiting on the system**

- **Processors have become so efficient that often the external CPI component is greater than the core component**
  - **So improvements in the chipset and bus efficiency can greatly improve system level performance (For processor intensive workloads)**
      - **Especially when the application has a high processor cache miss hit rates**
      - **High miss rates are often caused by large numbers of threads, large numbers of users and a very large data working set**
        - **The real world**

How We Improve Performance With

X3 Effectively Makes FSB Nearly Twice as Fast !!!

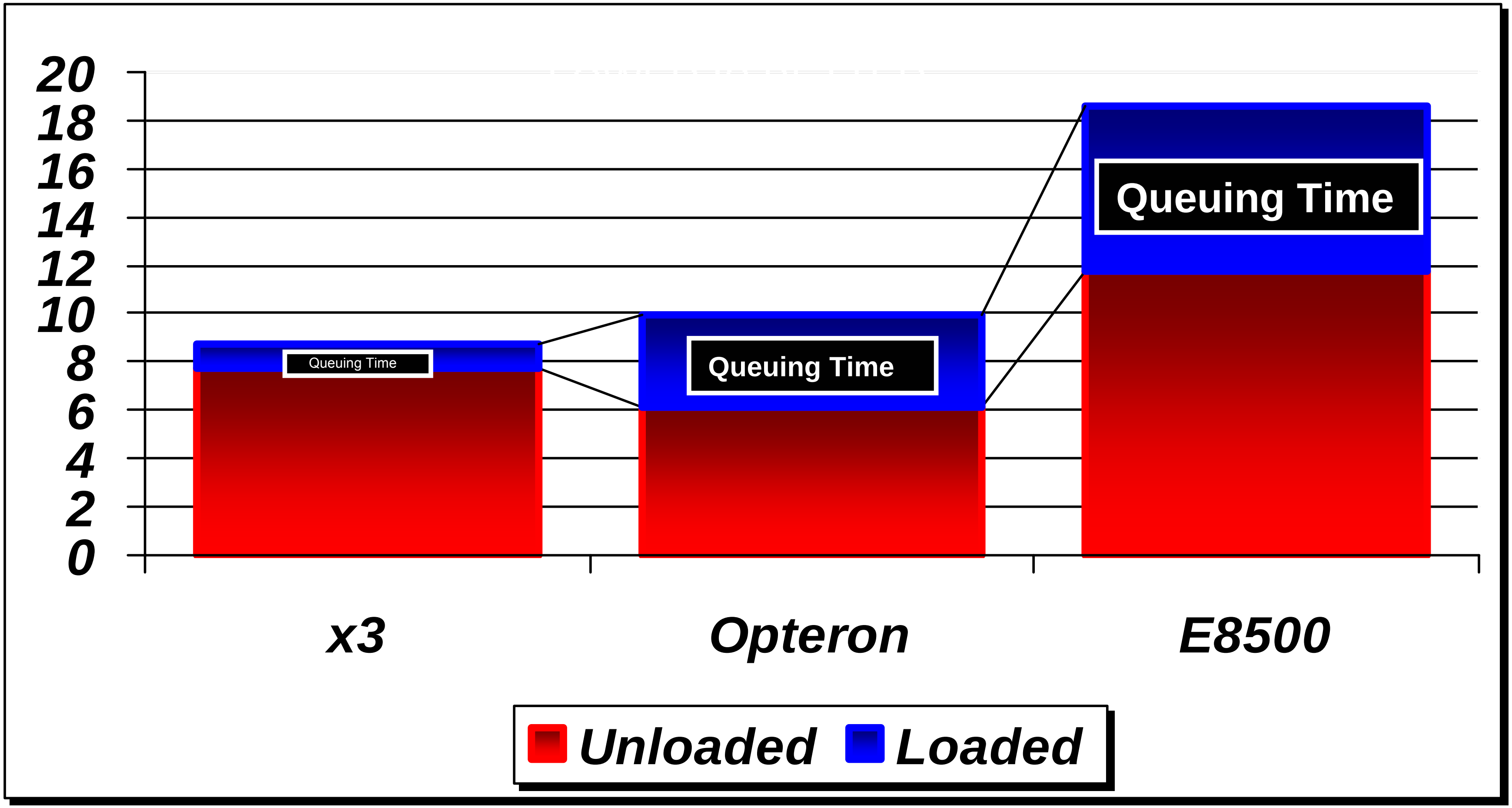


Snoop Filter Summary

- **Snoop filter is a method to improve commercial application performance**
  - **20 – 50% advantage compared to industry-standard systems without snoop filter**
  - **The faster the processor, the smaller the cache, the greater the gains from x3 snoop filter**
  - **With dual-core x3 performance analysis shows the x366 to be about 30 – 50% faster than same configuration Dell or HP dual-core model due to excessive loaded latencies associated with standard chipset queuing**
- **Only IBM is shipping snoop filter technology on x3 based systems**
  - **Commodity 4-way systems broadcast all addresses on twin front-side buses**
    - **Snoops must complete before memory can be accessed**
    - **All PCI traffic must reflect addresses on both front-side buses for snoop transactions before PCI memory read or write can complete**
    - **Increases memory access latency and queuing time…Net = Slower Performance!**
- **Expect to see snoop filters in future Intel and AMD Opteron components**

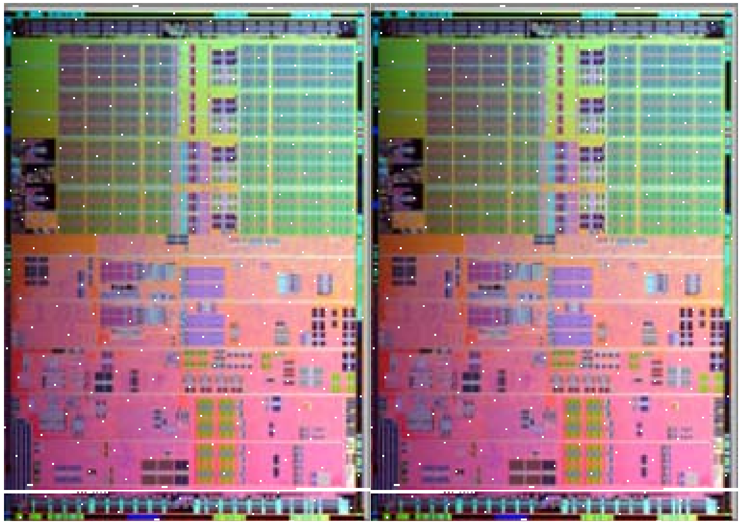
1) Loaded latency is the average latency of the chipset under load such as when running a heavy database workload
2) This is NOT related to the unloaded latency that manufactures like to brag about. Unloaded latency is MEANGINLESS!

It's Loaded Latency That Matters Everything Else is Marketing!

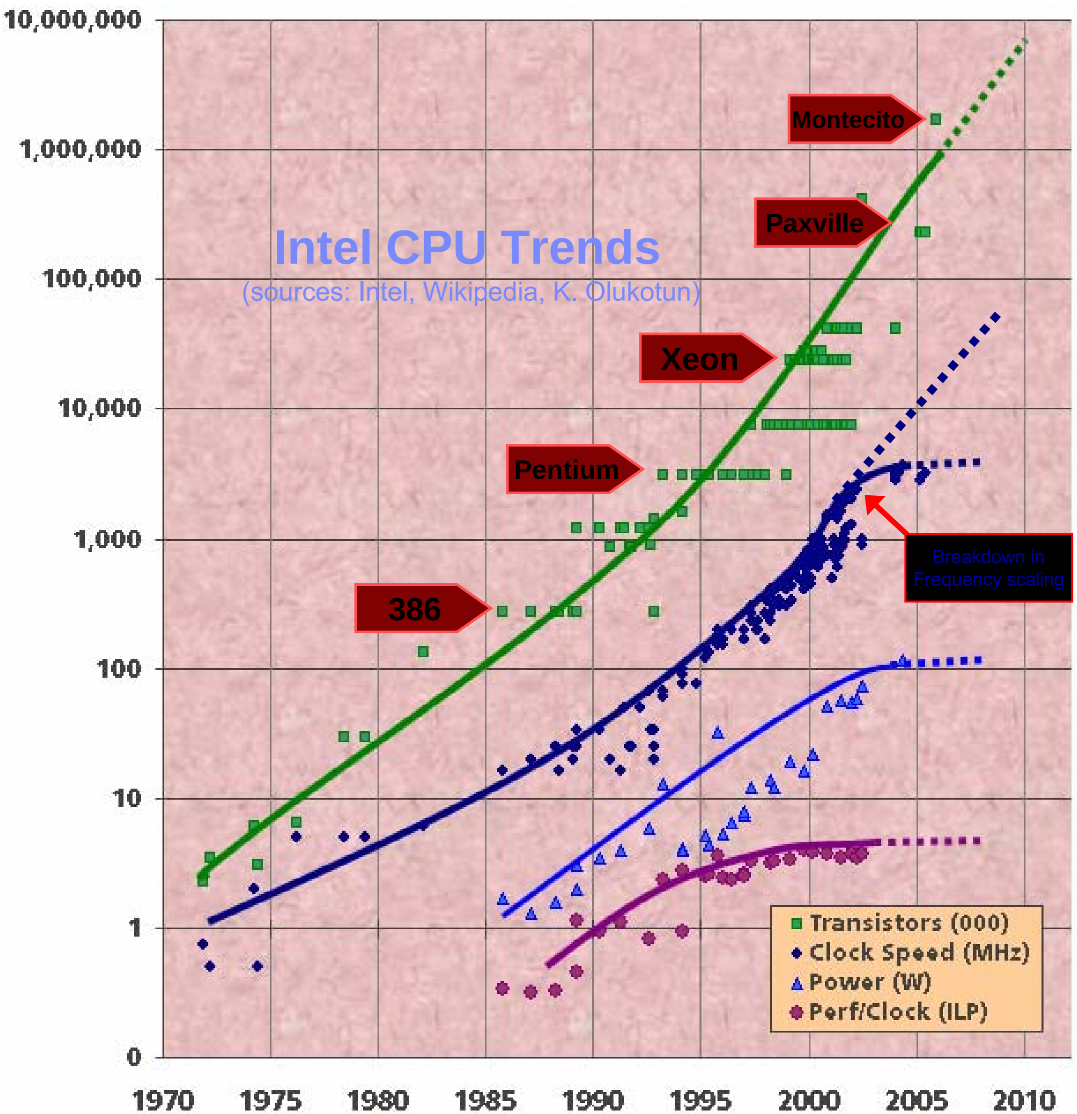


Memory latency for fastest processor configuration for each system
Loaded latency is average latency when running intensive database processing

Processor Performance / Issues & Futures



Technology Trend – Each Year We Get ~~Faster~~ More Processors



- **At around the beginning of 2003 physics began to limit processor frequency**
- **According to past trajectory we should be above 10GHz today!**
- **Historically, greater frequency increased single threaded and multi-threaded software performance without major code changes**
- **Multi-core will only improve software performance when an <u>increase</u> in the number of executing threads is possible**

Processor Futures

- Both AMD and Intel have significant processor architecture changes happening soon
- AMD – Rev F Processors
- Intel – Xeon DP - Core Micro-Architecture Processors
- Intel Xeon MP - Tulsa

AMD Rev F Overview

- **Current shipping versions of Opteron are Rev E**
  - **940 Socket**
- **Next generation is called Rev F (duh)**
  - **1207 Socket, so not socket compatible with current 940 pin Opteron**
  - **All Rev F Processors will be multi-core**
    - **No more single core Opteron**
  - **Support DDR2 Memory**
    - **8 DIMMs at 400MHz or 533MHz**
    - **4 DIMMs at 667MHz**
    - **Target of 2 DIMMs at 800MHz**
  - **Rev F retains existing HT1 (1GHz) HyperTransport technology**
  - **Rev F Adds PCI-Express**
  - **Rev F Adds RAS, Virtualization and Enhanced Power Mgt Technology**

AMD Server/Workstation Brand Positioning

PERFORMANCE 4-WAY AND 8-WAY

AMD Opteron™ 800 Series & 8000 Series Processors

– Designed for 4-way and 8-way Server solutions
– Only native x86 dual-core solution for 4-way / 8-way computing

PERFORMANCE 2-WAY

AMD Opteron™ 200 Series & 2000 Series Processors

– Designed for 2-way Server / Workstation solutions
– Only native x86 dual-core solution for 2-way computing

PERFORMANCE 1-WAY

AMD Opteron™ 100 Series & 1000 Series Processors

– Designed for 1-way Server / Workstation solutions
– Only native x86 dual-core solution for 1-way computing

AMD Opteron™ Processors for Servers and Workstations

- Direct Connect Architecture eliminates the bottlenecks inherent in front-side bus architectures by directly connecting CPU's, memory and I/O for reduced latency and optimized memory performance.
- Dual-Core AMD Opteron™ processors offer improved system efficiency and application performance
- AMD PowerNow!™ technology with Optimized Power Management decreases overall system power consumption without compromising system performance.

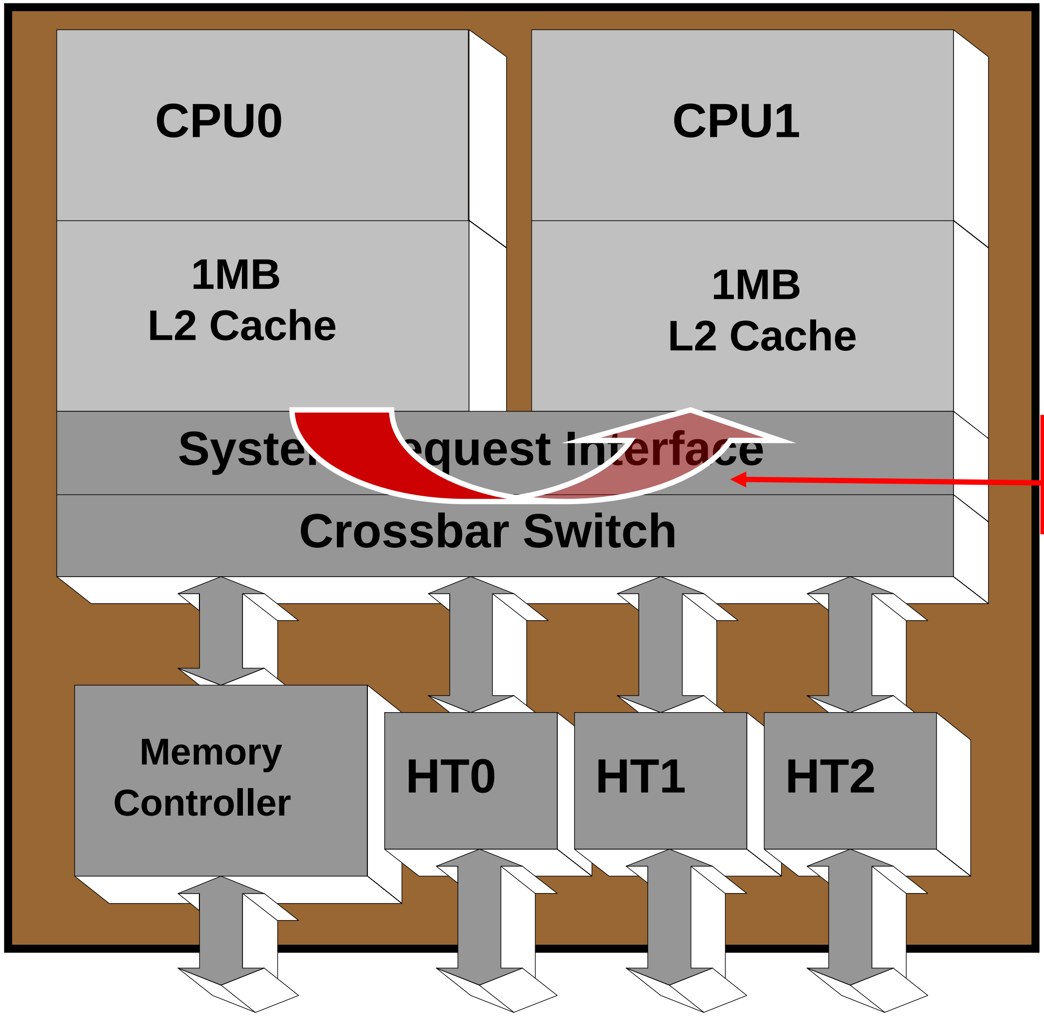
Source: AMD

AMD Opteron™ Processor Core Roadmap

	2H05	1H06	2H06
800 Series	**AMD Opteron™ Processor** Single & Dual-Core RDDR 1MB L2/core, SSE3 1GHz HT Socket 940	***“Santa Rosa“*** Dual-Core 1MB L2/core RDDR2, HT1 Memory RAS Security Virtualization Socket F (1207)	
200 Series	**AMD Opteron™ Processor** Single & Dual-Core RDDR 1MB L2/core, SSE3 1GHz HT Socket 940		
100 Series	**AMD Opteron™ Processor** Single/Dual-Core, UDDR 1GHz HT, SSE3 Socket 939		***“Santa Ana”*** Dual-Core 1MB L2/core UDDR2, HT1 Security Virtualization Socket AM2

Future | **Shipping** | **90nm SOI technology**

Opteron Rev E and F Dual-Core Design

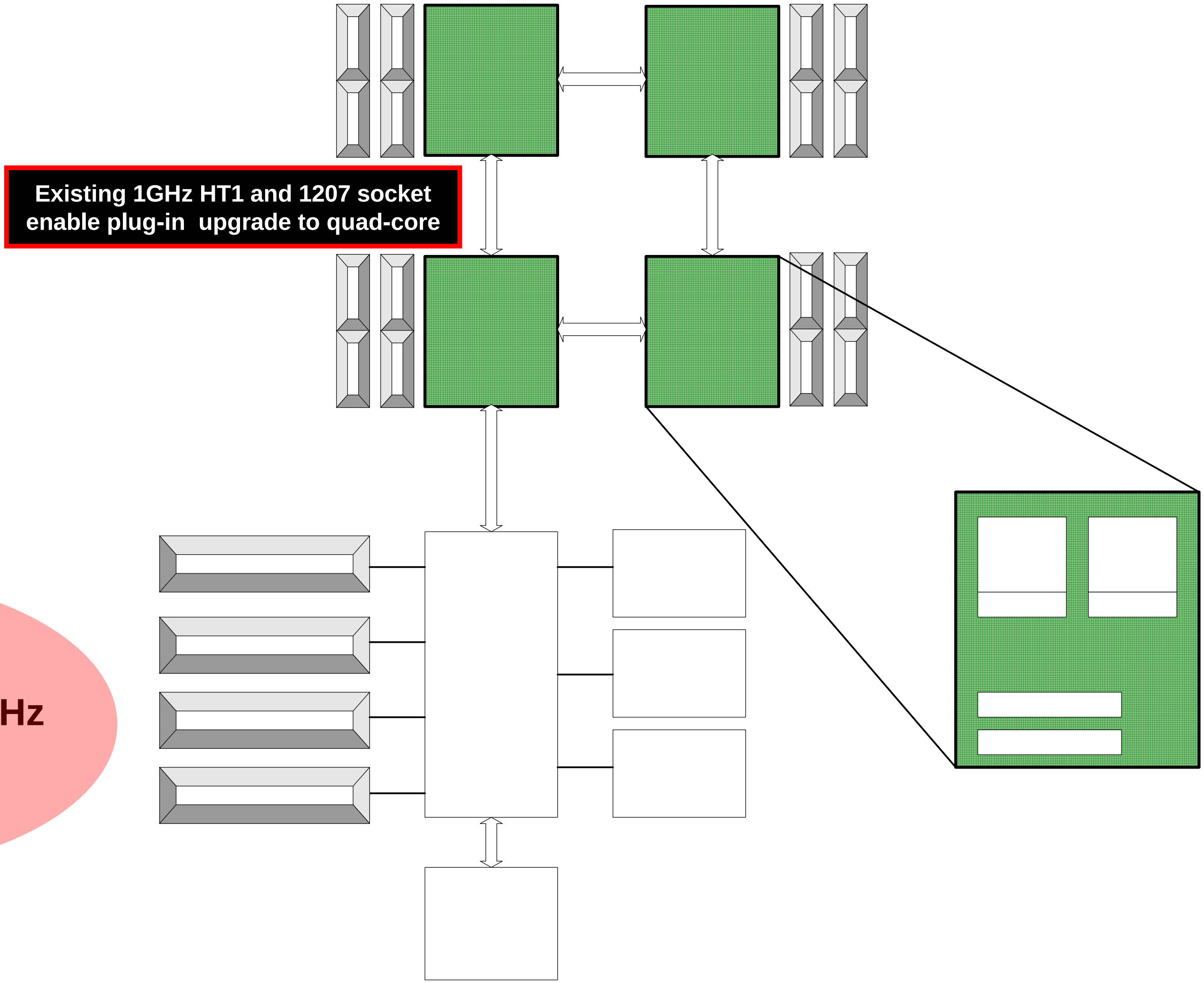


Cache to Cache data sharing is done through crossbar switch.

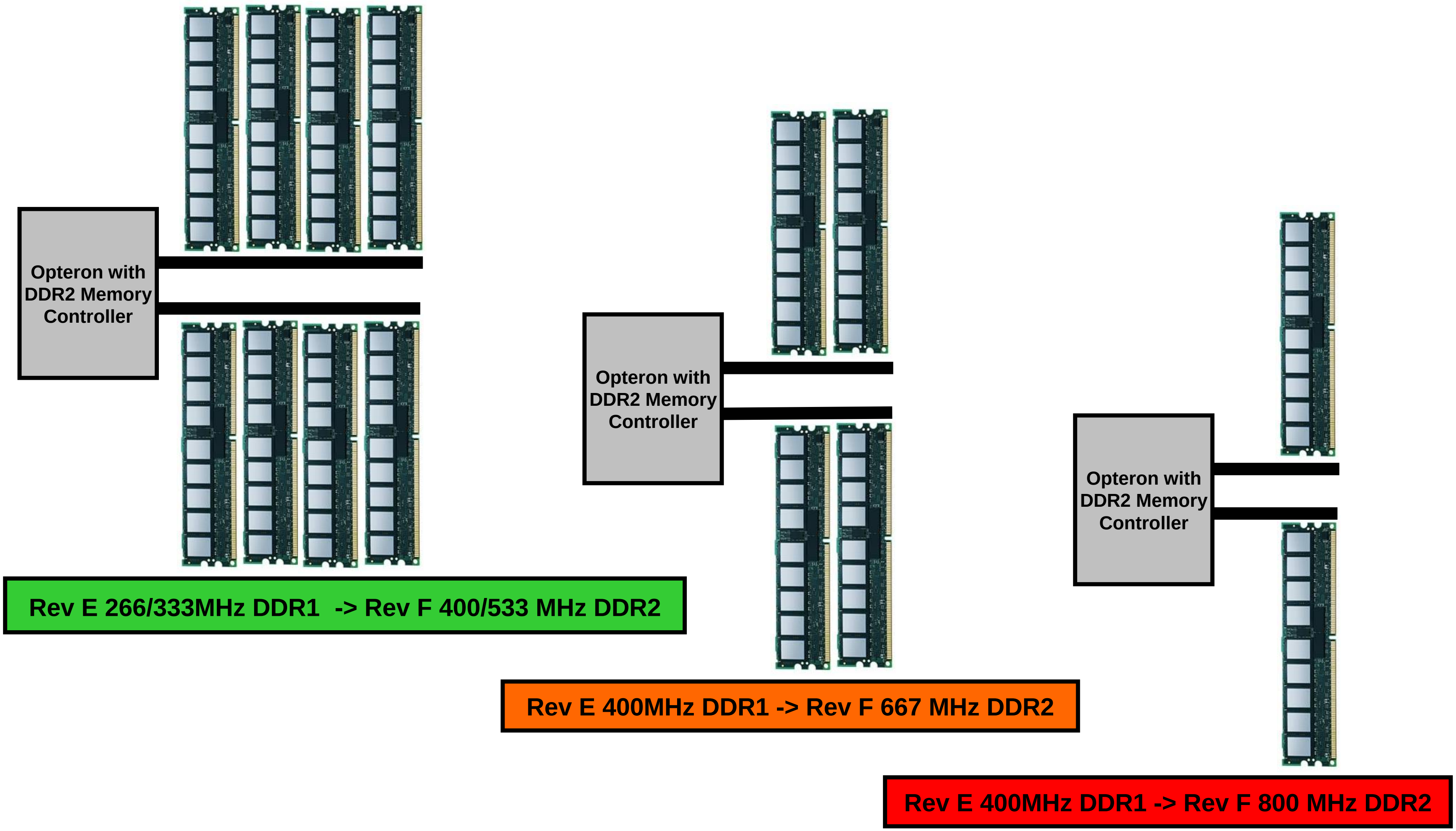
AMD Opteron™ Architecture

AMD Opteron™ Processor - Rev. F

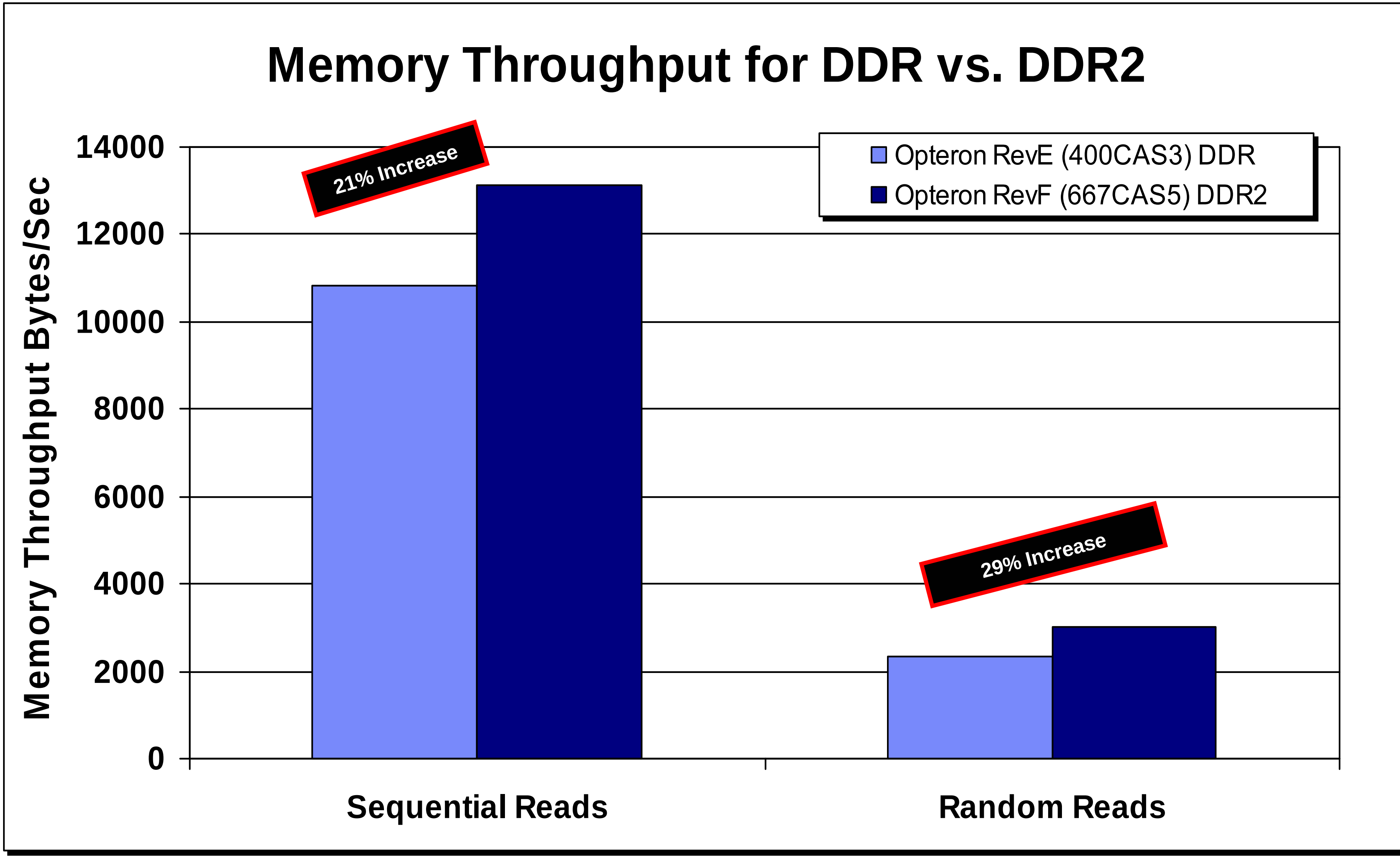
- Virtualization
  - Pacifica
- Security
  - Presidio
- Memory RAS
  - Online Spare
- Registered DDR-2
  - Launch at 667 MHz
  - Up to 800 MHz



Opteron to Add Faster DDR2 Memory Technology



DDR vs. DDR2 Memory Throughput



Intel Server Brand Positioning

Source: Intel

Q2 '06 | Q3 '06 | Q4 '06 | Q1 '07 | Q2 '07

Enterprise MP

Dual-Core Intel® Xeon® 7000 (Paxville MP)
667/800 FSB

Tulsa processor 7100 series 16MB L3
667/800FSB

Enabled chipsets -4P/8P+ platforms

IBM X3 chipset

Intel® Virtualization Technology

Intel® Cache Safe Technology (was codenamed Pellston)

Performance & Volume DP

Bensley Platform

Dual-Core Xeon® 5000
(Dempsey)
1066 FSB

Dual Core Intel® Xeon® Processor 5100
sequence (Woodcrest)
1066/1333 FSB

Clovertown

Intel® 5000P chipset
(Blackford)
ESB2/Gilgal(LAN)

Sunrise Lake IOP

Dual Core, FBD Memory,
Intel® VT, Intel® I/O AT,
Intel® ASM

Intel® Core™ Microarchitecture (Woodcrest)

Quad-Core

Value DP

Bensley-VS Platform

Dual-Core Xeon® 5000
(Dempsey)
667 FSB

Dual Core Intel® Xeon® Processor 5100 sequence
(Woodcrest)
1066 FSB

Intel® 5000V Chipset
(Blackford-VS)

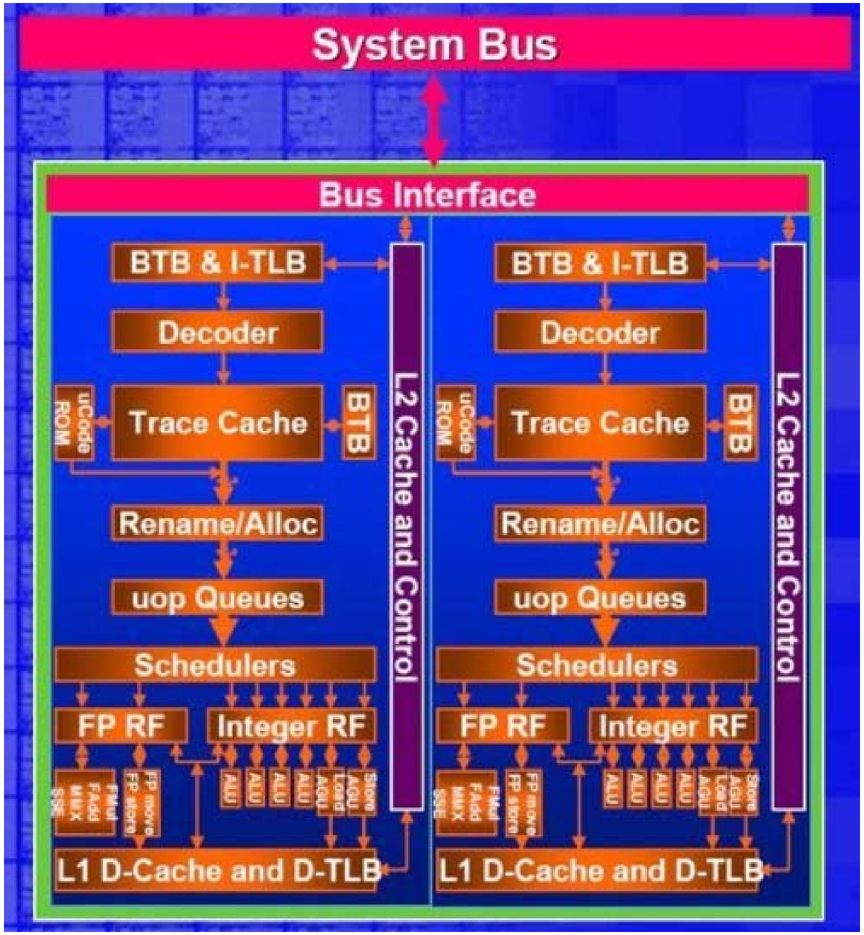
ESB2/Gilgal(LAN)

Sunrise Lake IOP

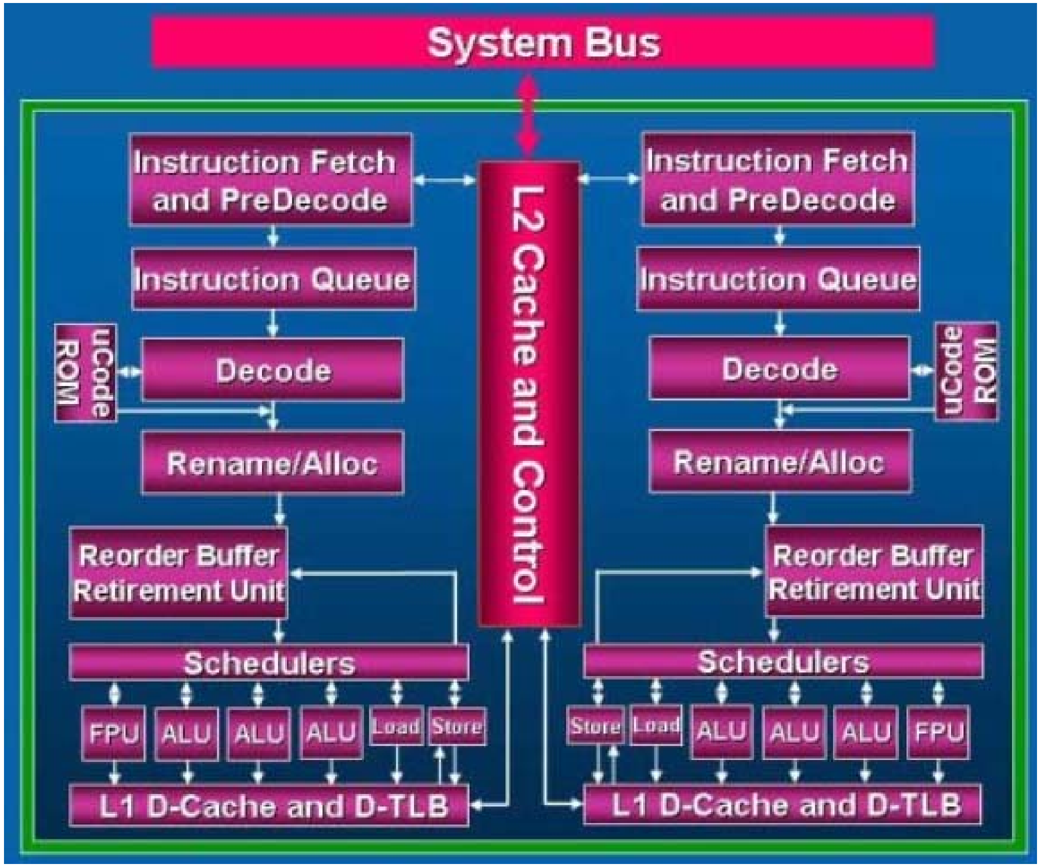
Dual Core, FBD Memory,
Intel® VT, Intel® I/O AT,
Intel® ASM

Intel® Core™ Microarchitecture (Woodcrest)

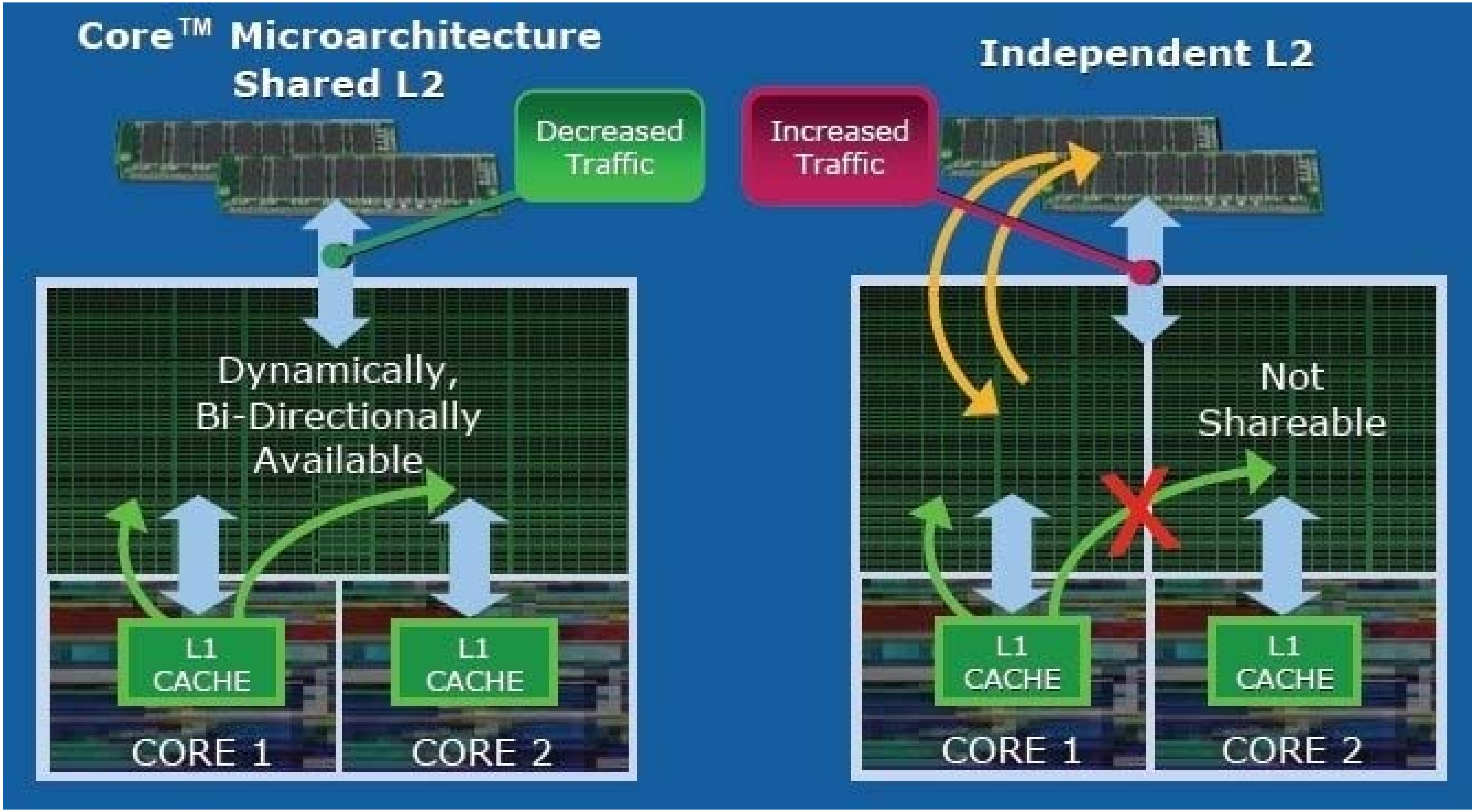
Intel‘s Paxville Processor Block Diagramm



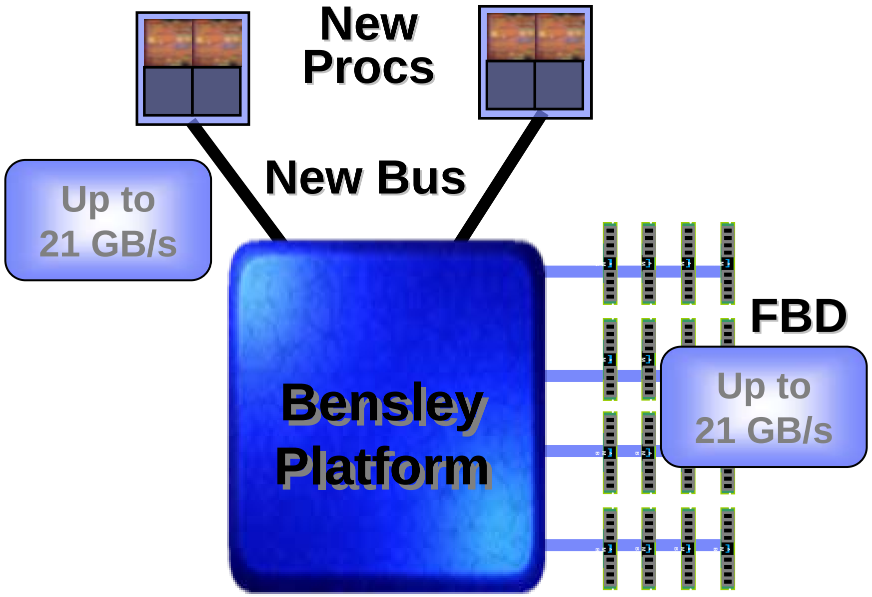
Intel‘s Conroe Processor Block Diagramm

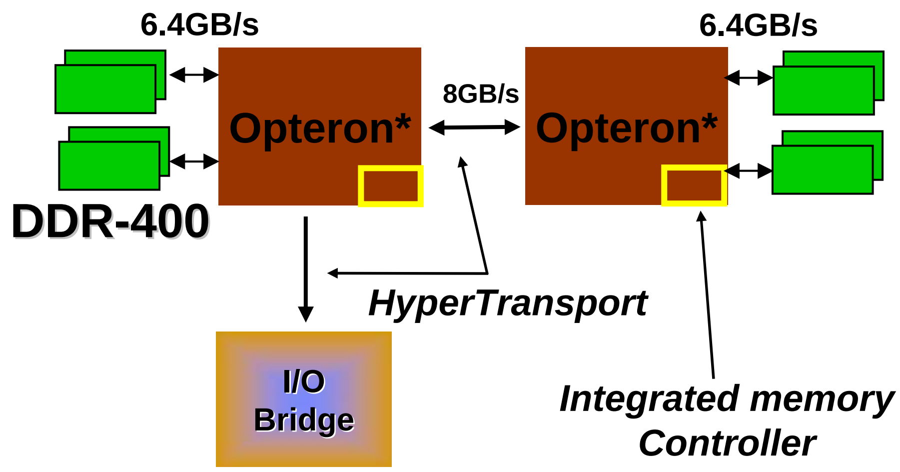


Intel's new architecture



2006: Intel Xeon DP vs. AMD Opteron



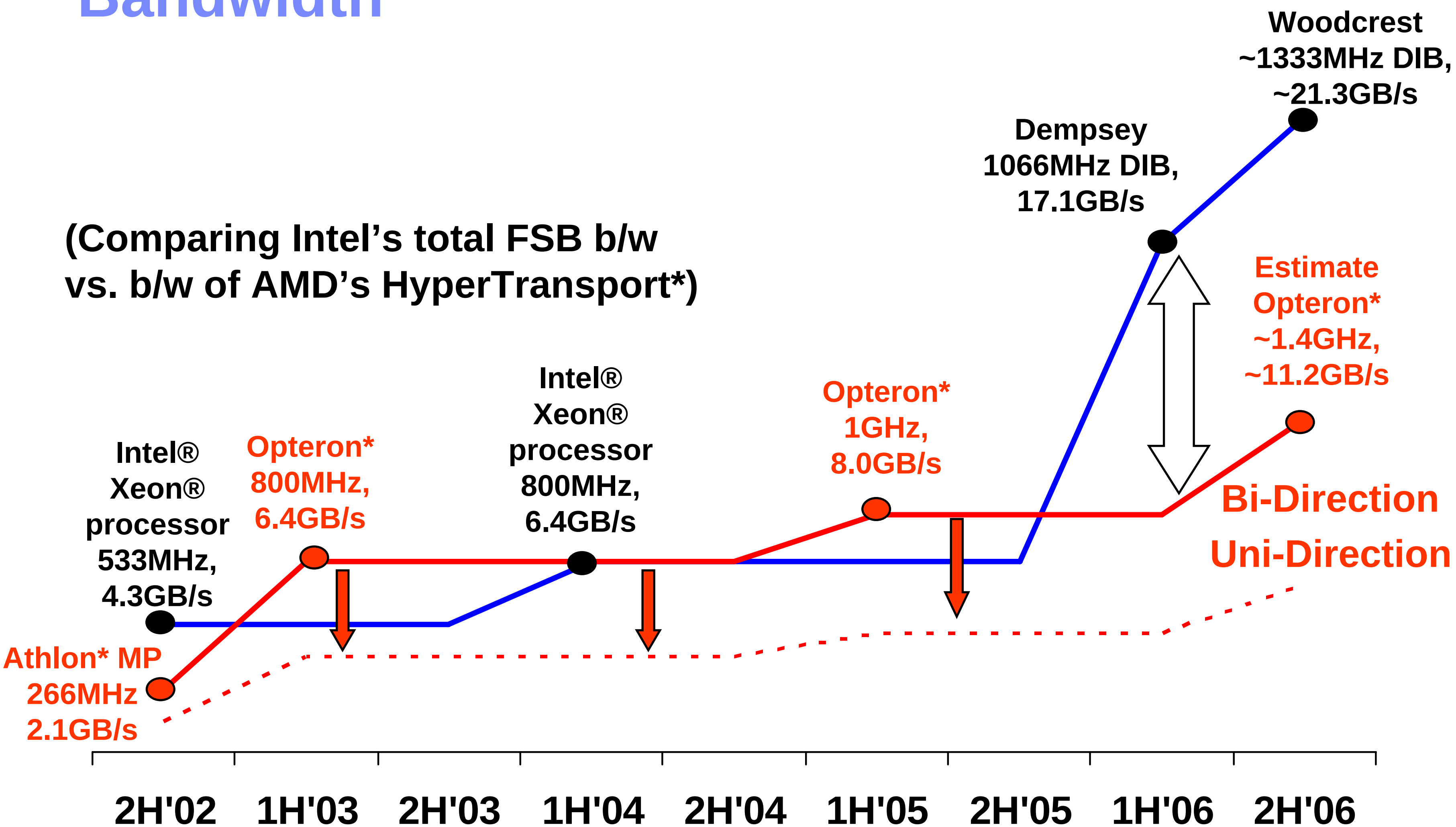


	Blackford / 2006	Opteron*
FSB BW peak	17 to 21 GB/s	8 GB/s
Memory BW peak	17 to 21 GB/s	12.8 GB/s*
Memory Capacity	64 GB	32-64 GB*

*** Mem b/w and capacity require both processor to be populated. Mem capacity is dependent on memory speed (Opteron requires trade-off of capacity for full freq & b/w)***

Front Side Bus vs. HyperTransport* Bandwidth

(Comparing Intel's total FSB b/w vs. b/w of AMD's HyperTransport*)

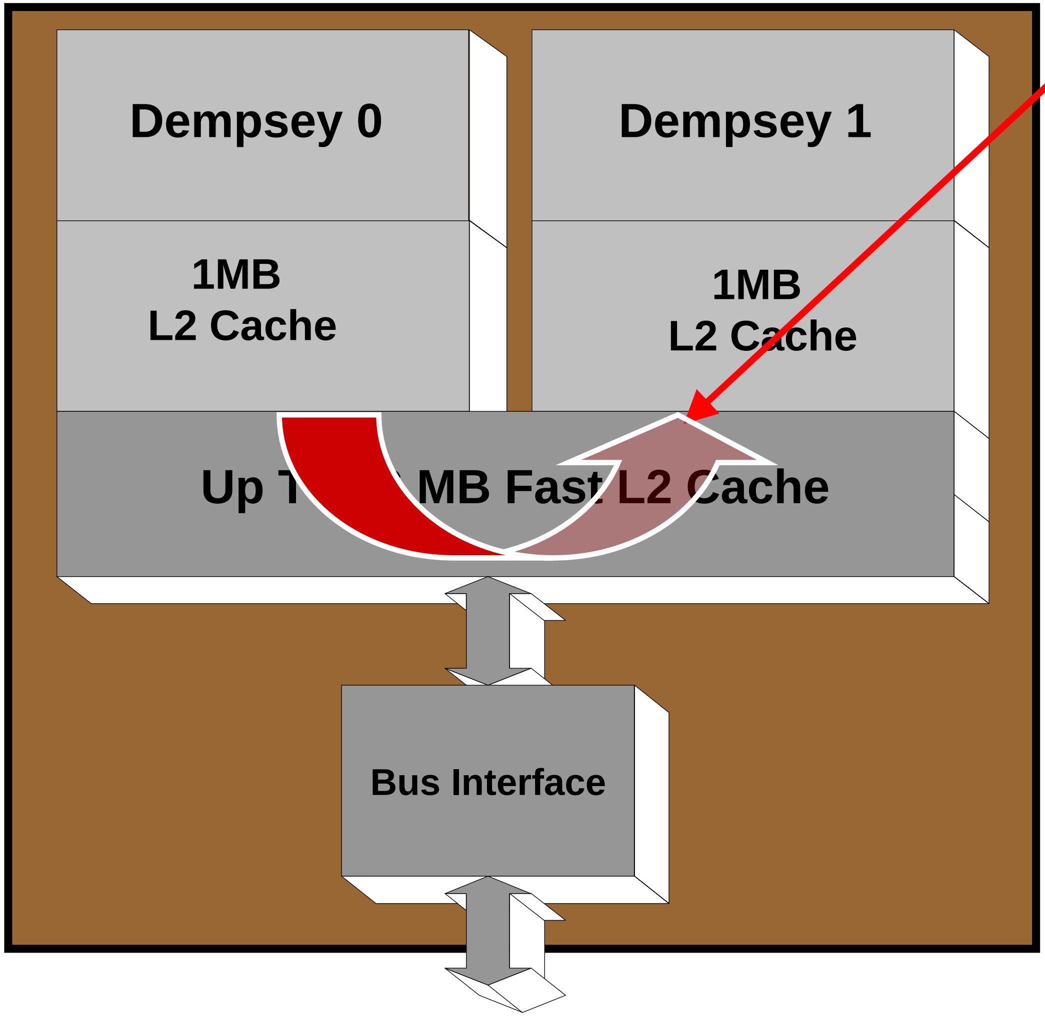


Source: Intel Corporation, as of Sept 26, 2005

All products, dates, comparisons and information are preliminary and subject to change without notice.

Intel Tulsa Architecture Overview

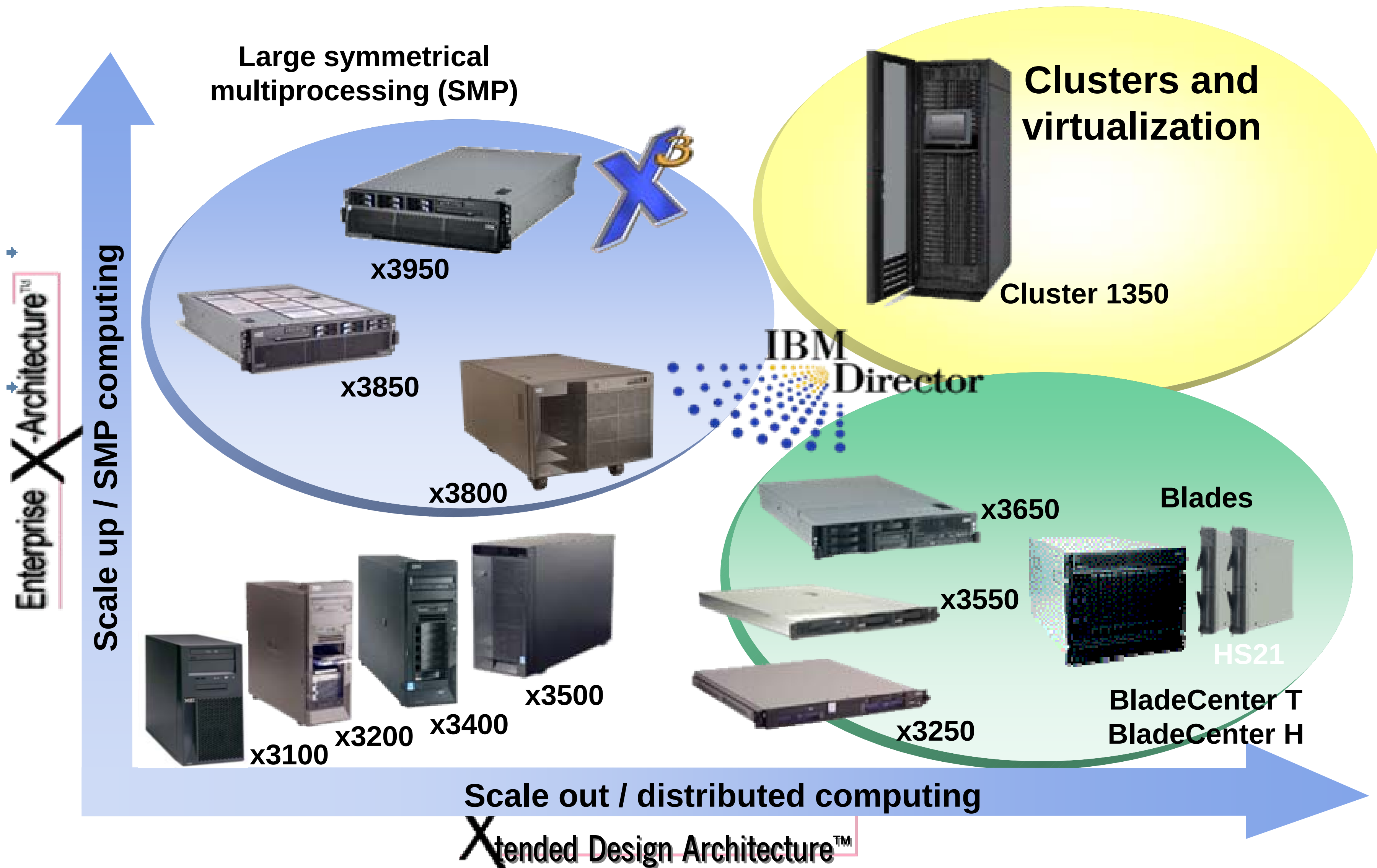
Cache to Cache data sharing is done through low latency shared L3 cache



Tulsa Architecture

- **16MB L3 3.4GHz Intro**
  - **Highest performance fast L3 cache**
    - **Much faster than Potomac L3 cache**
    - **Other frequencies and smaller cache sizes available**
  - **667 / 800MHz FSB**
  - **Hurricane 3 chipset will not support 800MHz**
    - **800MHz had a very minor performance gain on X3 because snoop filter reduces demand on FSB**
    - **But does have significant gain on TwinCastles because it does not have snoop filter and must run every cycle to FSB**

System x Portfolio is the Most Comprehensive in the x86 Industry



Thank you !

Anastasios Panagou
Advisory IT Specialist – FTSS System x & BladeCenter
IBM Schweiz
apa@ch.ibm.com

Trademarks and Notes

The following are trademarks of the International Business Machines Corporation in the United States and/or other countries. For a complete list of IBM Trademarks, see www.ibm.com/legal/copytrade.shtml: AS/400, DBE, e-business logo, ESCO, eServer, FICON, IBM, IBM Logo, iSeries, MVS, OS/390, pSeries, RS/6000, S/30, VM/ESA, VSE/ESA, Websphere, xSeries, z/OS, zSeries, z/VM

The following are trademarks or registered trademarks of other companies

Lotus, Notes, and Domino are trademarks or registered trademarks of Lotus Development Corporation
Java and all Java-related trademarks and logos are trademarks of Sun Microsystems, Inc., in the United States and other countries
LINUX is a registered trademark of Linux Torvalds
UNIX is a registered trademark of The Open Group in the United States and other countries.
Microsoft, Windows and Windows NT are registered trademarks of Microsoft Corporation.
SET and Secure Electronic Transaction are trademarks owned by SET Secure Electronic Transaction LLC.
Intel is a registered trademark of Intel Corporation
* All other products may be trademarks or registered trademarks of their respective companies.

NOTES:
Performance is in Internal Throughput Rate (ITR) ratio based on measurements and projections using standard IBM benchmarks in a controlled environment. The actual throughput that any user will experience will vary depending upon considerations such as the amount of multiprogramming in the user's job stream, the I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve throughput improvements equivalent to the performance ratios stated here.

IBM hardware products are manufactured from new parts, or new and serviceable used parts. Regardless, our warranty terms apply.

All customer examples cited or described in this presentation are presented as illustrations of the manner in which some customers have used IBM products and the results they may have achieved. Actual environmental costs and performance characteristics will vary depending on individual customer configurations and conditions.

This publication was produced in the United States. IBM may not offer the products, services or features discussed in this document in other countries, and the information may be subject to change without notice. Consult your local IBM business contact for information on the product or services available in your area.

All statements regarding IBM's future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only.

Information about non-IBM products is obtained from the manufacturers of those products or their published announcements. IBM has not tested those products and cannot confirm the performance, compatibility, or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

Prices subject to change without notice. Contact your IBM representative or Business Partner for the most current pricing in your geography.

References in this document to IBM products or services do not imply that IBM intends to make them available in every country. Any proposed use of claims in this presentation outside of the United States must be reviewed by local IBM country counsel prior to such use.

The information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this IBM product and use of those Web sites is at your own risk.